Master's Thesis of Philosophy

The Causal Exclusion Argument as an Argument Against Reductive Physicalism

환원적 물리주의에 대한 반대 논증으로서의 인과적 배제 논증

August 2023

Graduate School of Philosophy
Seoul National University
Western Philosophy Major

Eun Sung Choi

The Causal Exclusion Argument as an Argument Against Reductive Physicalism

Advisor: Sung-Il Han

Submitting a master's thesis of Philosophy

August, 2023

Graduate School of Philosophy
Seoul National University
Western Philosophy Major

Eun Sung Choi

Confirming the master's thesis written by
Eun Sung Choi
August 2023

Chair	김현섭	(Seal)
Vice Chair	한성일	(Seal)
Examiner	천현득	(Seal)

Abstract

The causal exclusion argument, in a nutshell, demonstrates that any physical event caused by a physically irreducible mental event is subject to causal overdetermination by another physical event. Jaegwon Kim argues that endorsing such systematic mental-physical causal overdetermination lacks motivation. Thus, it should be concluded that a purportedly physically irreducible mental cause is, in fact, either causally inefficacious or physically reducible. In other words, Kim contends that the lesson of the causal exclusion argument is that mentality must be reduced to save mental causation. He terms this position "conditional physical reductionism". Although conditional physical reductionism falls short of reductive physicalism *simpliciter*, it strongly aligns with reductive physicalism since relinquishing mental causation is highly undesirable.

However, many non-reductive physicalists object that the causal exclusion argument does not establish conditional physical reductionism. Instead, they advocate a strategy known as "causal compatibilism". According to this view, an effect can have more than one sufficient cause if there is a tight modal connection between the causes. Consequently, under causal compatibilism, systematic mental-physical causal overdetermination is unproblematic since the mental cause supervenes on the physical cause. While addressing this objection, Kim acknowledges that a "thick" conception of causation, referred to as the "productive conception of causation" is necessary to refute systematic mental-physical causal overdetermination. Consequently, Kim attempts to defend against causal compatibilism by advocating the productive conception of causation. The productive conception of causation posits that a cause is something that produces, or generates, or brings about its effects, something from which the effects *derive* their existence or occurrence.

Despite Kim's efforts, the question of whether the productive conception of causation is correct remains controversial. Nevertheless, both Kim and causal compatibilists concur that systematic mental–physical causal overdetermination is invalidated *if* the productive conception of causation is correct. This appears to suggest that the causal exclusion argument succeeds in establishing conditional physical reductionism *if* the productive conception of causation holds true.

The primary objective of this dissertation is to show that this is *not* the case. In other words, I argue that *even if* the productive conception of causation is correct, the causal exclusion argument does *not* substantiate conditional physical reductionism. This conclusion is reached by demonstrating that, within the framework of the productive conception of causation, mental properties can *solely* be reduced to causally inefficacious physical properties.

If my argument holds, it places Kim and other reductive physicalists in a dilemma. If Kim chooses to embrace the productive conception of causation, he will then need to acknowledge that reductive physicalism is merely a variant of epiphenomenalism. This consequence is unacceptable, given Kim's adherence to the reality of mental causation. Conversely, if Kim decides to reject the productive conception of causation, he will be compelled to embrace systematic mental–physical causal overdetermination, as there is no reason to justify its rejection. Once systematic mental–physical causal overdetermination is accepted, however, it is gratuitous to reduce mentality since it is possible to save mental causation using the causal compatibilist strategy. As a result, non–reductive physicalism becomes preferable to reductive physicalism. This consideration effectively demonstrates that the causal exclusion argument is, in fact, an argument *against*, not *for*, reductive physicalism.

Keywords : The Causal Exclusion Argument, Jaegwon Kim, Conditional Physical Reductionism, Reductive Physicalism, Causal Compatibilism, The Productive Conception of Causation
Student Number : 2020-28074

Table of Contents

Chapter 1. Introduction

Jaegwon Kim's causal exclusion argument (hereafter referred to as "the exclusion argument") is arguably one of the most famous and controversial arguments in the philosophy of mind. It demonstrates that the following five theses are mutually inconsistent:

> (Strong Supervenience) if any system *s* instantiates a mental property *M* at *t*, there necessarily exists a physical property *P* such that *s* instantiates *P* at *t*, and necessarily anything instantiating *P* at any time instantiates *M* at that time.
>
> (Distinctness) Mental properties are distinct from or are not identical with, physical properties.
>
> (Causal Efficacy) Mental properties have causal efficacy – that is, instantiations of a mental property *M* can, and do, cause other properties to be instantiated *in virtue of being an instance of M*.
>
> (Exclusion) No single event can have more than one sufficient cause occurring at any given time – unless it is a genuine case of causal overdetermination.
>
> (Closure) If a physical event has a cause that occurs at *t*, it has a sufficient physical cause that occurs at *t*.

The exclusion argument has two implications. Firstly, any mind-body theory that accepts all five premises must be rejected as incoherent. Secondly, any viable mind-body theory must deny at least one of the premises and provide a plausible explanation for doing so. Kim employs the first implication to argue *against* non-reductive physicalism. This is because, according to Kim, non-reductive

physicalism is committed to all five premises. Kim uses the second implication to argue *for* reductive physicalism, according to which (Distinctness) is false. This is because physicalists have only two options when confronted with the exclusion argument – either reject (Distinctness) or reject (Causal Efficacy). Kim calls this position "conditional physical reductionism":

> The position we have arrived at may be called *conditional physical reductionism*: the thesis that if mental properties are to be causally efficacious, they must be physically reducible. That is, to save mental causation we must reduce mentality. (Kim, 2005, 5)

Conditional physical reductionism strongly supports reductive physicalism because most philosophers do not regard epiphenomenalism or the denial of (Causal Efficacy) a genuine option.

However, there has been significant controversy regarding whether Kim was correct about the implications of the exclusion argument. Many non-reductive physicalists have contested that they are *not* obligated to accept all five premises of the exclusion argument, as they may reject (Exclusion). In the exclusion argument literature, this strategy of denying (Exclusion) is called "causal compatibilism" (or "compatibilism" for short). According to compatibilists, (Exclusion) is false because an effect can have more than one sufficient cause if there is a tight modal connection between the causes. The debate between compatibilists and Kim is illuminating because it revealed a crucial metaphysical assumption underlying (Exclusion). In the course of the dispute, it was acknowledged by both compatibilists and Kim that (Exclusion) is true *only if* a very robust or "thick" conception of causation, known as the "productive/generative conception of causation", is true:

> (The Productive/Generative Conception of Causation) a cause is something that produces, or generates, or brings

> about its effects, something from which the effects *derive* their existence or occurrence. (Kim, 2007, 235)

Therefore, the dispute regarding the truth of (Exclusion) naturally shifted into another dispute concerning the truth of the productive conception of causation. However, the question of whether the productive conception of causation is correct, or if some other "thin" conception of causation, such as the nomological or counterfactual conception of causation, can account for mental causation, is a contentious and challenging issue. Suffice it to say that Kim did not succeed in providing a satisfying justification of the productive conception of causation. This consideration shows that the exclusion argument fails to establish conditional physical reductionism, as (Exclusion) may be denied if the compatibilist can mount a persuasive objection against the productive conception of causation and provide a plausible account of mental causation under an alternative conception of causation.

All this will already be familiar to those who are acquainted with the exclusion argument literature. However, both compatibilists and Kim unreflectively assume that the exclusion argument does indeed establish conditional physical reductionism *if* (Exclusion) and the productive conception of causation are correct. In other words, as far as I know, no one doubts the truth of the position that may be called *weak conditional physical reductionism*: the thesis that if mental properties are to be causally efficacious *under the productive conception of causation*, they must be physically reducible. In this dissertation, I will depart from this convention by demonstrating that the exclusion argument does not even establish weak conditional physical reductionism. This will be done by arguing for the following claim: Given (Exclusion) and the productive conception of causation, denying (Distinctness) entails the denial of (Causal Efficacy). This implies that, under the productive conception of causation, reductive physicalism is a version of epiphenomenalism, rather than its rival.

If weak conditional physical reductionism is false, it becomes pointless for Kim and other reductive physicalists to defend (Exclusion) by advocating the productive conception of causation. In fact, the productive conception of causation is actually antithetical to their view, leading them straight to epiphenomenalism. However, reductive physicalists can neither seek to reject the productive conception of causation because it would imply the denial of (Exclusion). And this, in turn, would mean that non-reductive physicalism is a superior solution to the exclusion argument than reductive physicalism because non-reductive physicalism only denies (Exclusion) whereas reductive physicalism must deny both (Exclusion) and (Distinctness). Therefore, reductive physicalists find themselves in a dilemma. This consideration shows that the exclusion argument is actually an argument *against*, not *for*, reductive physicalism.

The structure of my dissertation can be summarized as follows: In chapter 2, I will explain the premises of the exclusion argument in detail. Next, I will expose the exclusion argument, as it is presented by Kim (2005). Then, I will examine the debate between non-reductive physicalists and Kim regarding the implications of the exclusion argument. By doing so, we will see that the productive conception of causation underlies (Exclusion). In chapter 3, I will present my main argument against weak conditional physical reductionism. I will first introduce its premises and then present the argument. Finally, I will respond to possible objections that may arise.

Chapter 2. The Causal Exclusion Argument for Conditional Reductive Physicalism

Until the 1960s, reductive physicalism was a widely accepted position among philosophers, which held that mental properties could be reduced to neural properties by way of identification. Suppose for instance that the mental property of [being in pain] correlates with the neurophysiological property of [C-fiber firing]. According to reductive physicalism, this correlation fact holds because [being in pain] *just is* [C-fiber firing]. However, things went rapidly downhill for this view when Hilary Putnam (1967) observed that it is empirically unlikely for mental properties to have a single neurophysiological correlate. It may be conceded that pain *in humans* correlates with [C-fiber firing]. Nonetheless, there are many non-human organisms, such as octopuses, mice, and dogs, also capable of feeling pain. And since humans and octopuses have heterogeneous neurophysiological structures, it is plausible that pain *in octopuses* will have a different neurophysiological correlate property, say, [O-fiber firing]. This consideration strongly supports the irreducibility of pain, as it prevents the identification of pain with any specific neurophysiological property. Therefore, in place of reduction, Putnam suggested that we regard [being in pain] as a multiply realized property with neurophysiological properties such as [C-fiber firing] and [O-fiber firing] among its realizers. This idea of multiple realization played a major role in establishing non-reductive physicalism as a prominent position in contemporary philosophy of the mind.

Kim devised the exclusion argument with the goal of challenging this prevailing orthodoxy regarding the mind-body relation. The argument takes the form of a *reductio ad absurdum*. Non-reductive physicalists must accept the following five premises: (Strong Supervenience), (Distinctness), (Causal

Efficacy), (Exclusion), and (Closure). However, these premises lead to a contradiction, making non-reductive physicalism an untenable position. As a result, reductive physicalism emerges as the only viable form of physicalism. However, the exclusion argument does not establish the truth of reductive physicalism *simpliciter*, as it only demonstrates that its premises are mutually inconsistent, not that any one of them is false. Nonetheless, Kim takes the exclusion argument to establish a weaker form of reductive physicalism he calls, "*conditional physical reductionism*: the thesis that if mental properties are to be causally efficacious, they must be physically reducible." (Kim, 2005, 5).

However, non-reductive physicalists did not easily concede that their position led to inconsistency and mounted various objections against the exclusion argument. This dispute has generated an immense literature, and subsequently illuminated the nature and limitations of the exclusion argument.

In this chapter, I will begin by introducing the premises of the exclusion argument, and then expose the argument as presented by Kim. Next, I will examine whether the exclusion argument succeeds in establishing conditional physical reductionism against non-reductive physicalism. Throughout this inquiry, I will uncover the metaphysical assumptions underlying the exclusion argument.

2.1. The Premises of the Causal Exclusion Argument

To begin the presentation of the exclusion argument, let us examine why Kim believes that non-reductive physicalism is committed to each premise. It is notoriously difficult to give an exact characterization of non-reductive physicalism because there is no consensus on the

notions of physicalism and reduction. However, for the purpose of presenting the exclusion argument, there is no need to choose one formulation of non-reductive physicalism at the expense of the others. Some rough sketches will suffice. Physicalism is the metaphysical thesis that everything is, in some sense, physical.[①] This does not necessarily imply that everything is *identical* to the physical. Otherwise, non-reductive physicalism would be conceptually impossible. Indeed, one of the most important conceptual challenges facing non-reductive physicalism is to capture the adequate determinative relation that must obtain between the physical and everything else if physicalism is to be true. Various candidate relations, such as supervenience, realization, and grounding have been proposed. Kim endorses supervenience-based formulations of physicalism.[②③]

---

① This is parallel to Thales' thesis that everything is water, and Berkeley's idealistic thesis that everything is mental.

② For those who are interested in other influential formulations of physicalism, see Daniel Stoljar's entry "Physicalism" in the *Stanford Encyclopedia of Philosophy*.

③ It is an interesting exegetical issue whether Kim was committed to supervenience-based formulations of physicalism throughout his entire career. In some of his writings, Kim himself acknowledges that supervenience may not be adequate as a relation of metaphysical determination:

> The (...) problem I have in mind concerns the question whether [supervenience] is a genuine relation with explanatory force and metaphysical significance. (...) Supervenience, as it is standardly understood (...), does not represent a unitary relation of metaphysical or explanatory interest and significance. Supervenience can obtain for all sorts of reasons. [Why does mind-body supervenience obtain?] (...) A (...) divergent range of explanations has been offered: (1) mental phenomena are caused by physical phenomena; (2) mental properties are definable in terms of behavioural/physical properties; (3) mind and body are simply two aspects of some deeper reality that is neither mental nor physical in itself; and, of course, (4) the

The intuitive notion of supervenience is captured by the following slogan: "If *F* supervenes on *G*, there is no *F*-difference without *G*-difference". Therefore, when one says that the mental supervenes on the physical, she is saying that two entities which are alike in physical respects cannot differ in mental respects. This idea can be regimented in various ways. One widely accepted version of supervenience physicalism presented by Frank Jackson invokes global supervenience, which refers to world-wide patterns of supervenience:

> mental emerges from the physical. Since supervenience is consistent with all of these relations, it cannot in itself be a single homogeneous relation. Supervenience simply states an interesting pattern of co-variation between two sets of properties, the normative and the non-normative, the mental and the physical, and so on. (Kim, 2006, 200)

However, in other works, Kim maintains that supervenience is a relation of generation or determination:

> What I have in mind is (..) the fundamental notion of (...) determination. Another way in which a state, or property instance, is generated is supervenience; the aesthetic properties of a work of art are generated in this sense I have in mind by its physical properties. So are moral properties of acts and persons generated by their nonmoral, descriptive properties. It is the relation that sanctions the assertion that something has a certain property because, or in virtue of the fact that, it has certain other properties that generate it. (Kim, 2005, 18)

Especially, in his 2009 paper "From Naturalism to Physicalism: Supervenience Redux", Kim attempts to defend supervenience as "a way of thinking about the dependence of the mental on the physical" (Kim, 2009, 122). Therefore, it seems that Kim remained faithful to supervenience-based formulations of physicalism until the end of his career.

> (Global Supervenience) Any world which is a minimal physical duplicate of our world is a duplicate *simpliciter.* (Jackson, 1998, 12)

Here, a minimal physical duplicate of our world is what results from duplicating *all and only* the physical facts that obtain in our world. Although (Global Supervenience) is endorsed by physicalists of all stripes, many believe that a stronger supervenience relation must hold between the physical and everything else. They feel the need to tell us *why* such a world-wide pattern of supervenience obtains between physical and non-physical facts. Here is a plausible answer: the world-wide pattern of supervenience obtains *because* each non-physical property has some physical property as its supervenience base. This idea is captured by the following stronger supervenience claim:

> (Strong Supervenience) if any system s instantiates a mental property M at t, there necessarily exists a physical property P such that s instantiates P at t, and necessarily anything instantiating P at any time instantiates M at that time. (Kim, 2005, 33)

Now, it is true that supervenience-based formulations of physicalism have faced strong criticisms from various authors, such as Terence Horgan (1993), Andrew Melnyk (2003), and Jessica Wilson (2005). These authors pointed out that supervenience is not a determinative relation, but rather only tracks an interesting pattern of co-variance between two sets of properties. As a result, mind-body supervenience can be embraced by various dualist positions, including parallelism and epiphenomenalism. Nonetheless, these objections have to do with the fact that supervenience is too weak to characterize physicalism, not that it is too strong. Therefore, even critics of supervenience-based formulations of physicalism can recognize mind-body supervenience as a necessary condition for physicalism.

Let us now turn to reduction. Reduction is another concept whose exact meaning is not yet fixed. The root meaning of "reduction" in philosophical contexts was first given by J.J.C. Smart (1959, 170) when he said that sensations are "nothing over and above" brain processes. More generally, if *x* reduces to *y*, then *x* is nothing over and above *y*. Now, various models of reduction indicate that reduction is *not* the same as identity. For example, Ernest Nagel's (1961) model of theory reduction, which has been influential in reductionism literature, states that a theory *TR* reduces to another theory *TB* if and only if the laws of *TR* can be derived from the laws of *TB* with the possible help of relevant bridge laws or coordinating definitions. To elaborate, suppose that psychology includes the psychophysical law $M \rightarrow P*$ as one of its laws. According to Nagel's model, reduction of psychology to physics will involve the derivation of $M \rightarrow P*$ from the physical laws and bridge laws. Suppose that the physical law relevant to this task is $P \rightarrow P*$. Then, the bridge laws can be regarded as identities of the form $M = P$, but it is usually supposed that they can also be seen as *nomic equivalences* of the form $M \leftrightarrow P$, as nomic equivalences are sufficient for deriving the target laws. For example, $P \rightarrow P*$ and $M \leftrightarrow P$ entail $M \rightarrow P*$ in a straightforward way.

Now, even though reducibility does not imply identity, it must be acknowledged that *irreducibility* implies non-identity or distinctness. For how could M be identical with P and yet be ontologically irreducible to P? Consequently, by asserting that mental properties are irreducible to physical properties, non-reductive physicalists are thereby accepting the following premise:

> (Distinctness) Mental properties are distinct from or are not identical with, physical properties.

So far, we have identified two premises endorsed by non-reductive physicalism that follow from its commitment to physicalism and anti-reductionism. According to Kim, there are three more premises that

any serious non-reductive physicalist must accept. The third premise concerns the causal status of physically irreducible mental properties:

> (Causal Efficacy) Mental properties have causal efficacy. (Kim, 2005, 35)

According to the standard accounts of event causation, it is events that enter into causal relations, rather than properties themselves. So what does it mean for a property to be causally efficacious? One might suggest that the causal efficacy of properties is *nothing over and above* the causal efficacy of their instantiations, in the following sense:

> (Property Efficacy*) a property *P* is causally efficacious if and only if instantiations of *P* can, and do, cause other properties to be instantiated.

However, (Property Efficacy*) is unsatisfactory because while the causal efficacy of *P*-instantiations or *P*-events is *necessary* for the causal efficacy of *P*, it is by no means *sufficient*. To understand this point, we need to briefly delve into the debate surrounding Donald Davidson (1970)'s famous causal argument for physicalism, known as "anomalous monism".[④] Davidson's argument begins by proposing the following three premises:

> (a) *Nomological character of causality*: if an event *c* causes another event *e*, there is a *strict law* that subsumes *c* and *e*.
>
> (b) *Mind-body causation*: some mental events cause, and are caused by, physical events.
>
> (c) *Anomalism of the mental*: there are no *strict laws* governing mental phenomena. In other words, there are

④ I found Kim's (2007; 2009) exposition of Davidson's anomalous monism and the critiques of this position to be valuable.

neither strict psychological laws nor strict psychophysical laws.

What are strict laws? *Strict* laws, unlike *ceteris paribus* laws, are exceptionless laws that are not hedged by ceteris paribus clauses. Further, laws, unlike mere empirical generalizations, are true generalizations that support causal counterfactuals and are confirmable by observations of positive instances (that is, they are inductively projectible). According to Davidson, strict laws, if they exist, are to be found only in developed physics. Now, at first glance, (a), (b), and (c) appear to be mutually inconsistent. According to (b), there are cases of mental-to-physical causation. So let us suppose that a mental event m causes a physical event p^*. Then, according to (a), there must be a strict law L that subsumes m and p^*. But this seems to contradict (c), as it dictates that there are no strict psychophysical laws. However, Davidson astutely replies that there is no contradiction if L is a physical law. In other words, the causal efficacy of m is not threatened as long as m possesses physical properties, thereby counting as a physical event.[5] This reasoning can be generalized to establish (d):

[5] To be precise, this strategy for securing the causal efficacy of mental events is unavailable for those who subscribe to a fine-grained conception of events, where events involve *only one* property. According to Kim's view, events are property instantiations, i.e., instantiations *of* properties *by* objects *at* times. Kimean events are fine-grained because an event that is an instantiation of one property cannot simultaneously be an instantiation of another property unless the two properties are identical. Coarse-grained conceptions of events, on the other hand, allow an event to involve multiple properties. According to Davidson, "an event is mental (or physical) just in case a mental (or physical) description, or predicate, is true of it—or, as we might say, in case it falls under a mental (or physical) kind" (Kim, 2009, 36). Davidsonian events are coarse-grained because there is no issue with an event satisfying both mental and physical descriptions. In this essay, I will assume that events are fine-grained entities in Kim's sense.

(d) *Physical Monism*: every causally interacting mental event is identical to some physical event.

Notwithstanding its ingenuity, many philosophers (such as Kim, 1984; McLaughlin, 1989) have objected that anomalous monism fails to provide a satisfying account of mental causation. The common criticism is that Davidson's theory does not give any causal role to mental properties or kinds. To elaborate, according to anomalous monism, a mental event *m* is causally efficacious *only because* it is subsumed under a strict physical law. And an event is subsumed under a strict physical law *solely in virtue of* its physical properties. As a result, *m*'s causal relations are fully and exclusively determined by its physical properties, implying that *m*'s mental properties are causally irrelevant. Because of this, anomalous monism is generally perceived as a form of mental *type* or *property* epiphenomenalism, although it is not a form of mental *token* epiphenomenalism (this distinction is due to Brian McLaughlin, 1989).

Now, it is a controversial matter whether *Davidson's* anomalous monism leads to property epiphenomenalism. This is because Davidson was a nominalist rather than a realist about properties. However, regardless of Davidson's view about properties, Kim and McLaughlin's objection demonstrates that anomalous monism, when combined with property realism, results in property epiphenomenalism. And of course, we are assuming property realism here, as we are talking about the causal efficacy of properties. Therefore, without going further into this issue, let us suppose that the critics raised an adequate objection to anomalous monism (under property realism). Given such an assumption, the lesson is that even under anomalous monism, instantiations of mental properties (i.e., mental events) can, and do, cause other properties to be instantiated. This constitutes a counterexample to (Property Efficacy*) because anomalous monism fails to secure the causal efficacy of mental properties. Therefore, I propose the following revised necessary and sufficient condition for property efficacy:

> (Property Efficacy) a property *P* is causally efficacious if and only if instantiations of *P* can, and do, cause other properties to be instantiated *in virtue of being an instance of P*.⑥

Note that anomalous monism does not provide an objection to (Property Efficacy), as the view posits that mental events cause other events *solely in virtue of* being an instance of a physical property.

With the causal efficacy of properties understood as stated in (Property Efficacy), most non-reductive physicalists will agree that (Causal Efficacy) is non-negotiable. Jerry Fodor, one of the most prominent non-reductive physicalists, voices the following concern over the possible loss of mental causation:

> If it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for saying (...), if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world. (Fodor, 1989, 77)

Indeed, non-reductive physicalists are so eager to prove (Distinctness) because they believe that mental properties can make *distinct contributions* to the way the world unfolds. As regards epiphenomenalism, the position that denies (Causal Efficacy), Kim (1999, 22) agrees with Samuel Alexander's observation that it is not substantially different from eliminativism:

---

⑥ This condition is endorsed by Kim (2005, 42) when he says, ""An *M*-instance causes a *P*-instance" must be understood with the proviso "in virtue of the former being an instance of *M* and the latter an instance of *P*.""

> Epiphenomenalism supposes something to exist in nature which has nothing to do, no purpose to serve, a species of *noblesse* which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time be abolished. (Alexander, 1920, 8)

The fourth and final premises of the exclusion argument are claims about the causal structure of the world:

> (Exclusion) If an event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e* (unless this is a genuine case of causal overdetermination). (Kim, 2005, 17)

By "a genuine case of causal overdetermination", we are referring to the textbook cases of causal overdetermination like two bullets hitting the victim's heart at the same time. Such cases involve two or more *separate* and *independent* causal chains converging at a common effect. Consequently, each overdetermining cause plays a *distinct* and *distinctive* causal role.

However, unlike the other premises, it is not immediately obvious why non-reductive physicalists should endorse (Exclusion). In fact, many non-reductive physicalists have responded to the exclusion argument by attempting to refute this premise. Since the debate surrounding (Exclusion) is crucial to fully understanding the nature of the exclusion argument, we will revisit it in section 2.3.2. after presenting the argument itself. For now, let us move on.

> (Closure) If a physical event has a cause that occurs at *t*, it has a sufficient physical cause that occurs at *t*.[⑦]

[⑦] There are two things to note about the formulation of (Closure) used in this text. First, Kim's formulation of (Closure) is slightly different. He does not say that the physical cause that occurs at *t* is a *sufficient* cause:

The intuitive idea behind (Closure) is that the physical realm is causally self-sufficient, which means that any physical effect can be causally accounted for without reference to some non-physical entity. It widely accepted that physicalism is intimately related to (Closure). Bennett (2008, 282) claims that "physicalism itself arguably entails it". According to David Papineau (2001), physicalism rose to prominence in the 20th century largely because scientific investigations during that period established the empirical plausibility of (Closure). Nevertheless, (Closure) is not exclusively a physicalist thesis. It is compatible with forms of mind-body dualism that reject mind-body interaction, such as Spinoza's parallelism or Leibniz's pre-established harmony account.

> (Closure) If a physical event has a cause that occurs at *t*, it has a physical cause that occurs at *t*. (Kim, 2005, 43)

However, this formulation is inadequate because it allows a physical event to be jointly caused by a mental event occurring at *t* and a physical event occurring at *t*. This consequence fails to reflect the guiding idea behind (Closure), which is that the physical domain is causally self-sufficient. Furthermore, the exclusion argument wouldn't get off the ground because the mental cause occurring at *t* and the physical cause occurring at *t* would not causally overdetermine their physical effect. For these reasons, Loewer (2007, 251) also adopts the version of (Closure) used in the text.
Second, this version of (Closure) fails to account for cases of indeterministic causation. However, this problem can be easily accommodated by adopting the following formulation:

> (Closure) The objective probability of every physical event is fixed by prior physical facts and laws alone (O'Connor and Wong, 2005, 658)

However, since the possibility of indeterministic causation is not crucial in presenting the exclusion argument, we will stick to the simpler formulation of (Closure) used in the text. Readers who are interested in various formulations of (Closure) can refer to Lowe (2003).

This is because (Closure) does not claim that the world consists solely of physical entities or events, or that physical causation is the only kind of causation that occurs. Its only implication is that non-physical events cannot have independent causal influence on physical events.

2.2. The Causal Exclusion Argument

With the requisite premises established, we can now examine the exclusion argument. As Kim (2005) suggests, we break the argument down into two stages. In the first stage, we will see that, given mind-body supervenience, any instance of mental-to-mental causation entails an instance of mental-to-physical causation. This implies that, if mental-to-physical causation is eliminated, all mental causation must also be eliminated under the physicalist worldview. In the second stage, we will demonstrate that accepting all five premises of the exclusion argument renders mental-to-physical causation incoherent. By combining the two stages, we can see that rejecting at least one of the five premises, except (Causal Efficacy), is necessary to preserve mental causation.

Before proceeding with the argument, I will once again lay out the premises to make the structure of the argument more conspicuous:

> (Strong Supervenience) if any system s instantiates a mental property M at t, there necessarily exists a physical property P such that s instantiates P at t, and necessarily anything instantiating P at any time instantiates M at that time.
>
> (Distinctness) Mental properties are distinct from or are not identical with, physical properties.
>
> (Causal Efficacy) Mental properties have causal efficacy – that is, instantiations of a mental property M can, and do,

cause other properties to be instantiated *in virtue of being an instance of M*.

(Exclusion) No single event can have more than one sufficient cause occurring at any given time – unless it is a genuine case of causal overdetermination.

(Closure) If a physical event has a cause that occurs at *t*, it has a sufficient physical cause that occurs at *t*.

Stage 1

According to (Causal Efficacy), there are cases of mental causation. And any case of mental causation is either a case of mental-to-mental causation or a case of mental-to-physical causation. Let us first deal with cases of mental-to-mental causation. Let *M* and *M** be mental properties. Let *m* be an event which is an instantiation of *M* at time *t*. Let *m** be an event which is an instantiation of *M** at time *t**. Now suppose that:

(1) *m* causes *m**.

Now, (Strong supervenience) dictates that *M** has a physical property *P** as its supervenience base. Hence, there is a physical event *p** which is an instantiation of *P** at time *t** such that:

(2) *m** has *p** as its supervenience base.

But there is a tension between (1) and (2) because both *m* and *p** are sufficient for *m**. To elaborate, given *p**, it will necessitate *m** no matter what happened before *t**. This means that *m** would have occurred even if *m* had not. Thus, *m*'s status as the cause of *m** is undermined. To resolve this tension, we need to assume that the instantiation of *m* is somehow responsible for the instantiation of *p**.

But how? By causing *p**. Hence, we get the following case of mental-to-physical causation:

(3) *m* causes *p**.

Thus, we get the result that any case of mental-to-mental causation presupposes a case of mental-to-physical causation. This completes the first stage of the exclusion argument.

Stage 2

We will start from (3), a putative case of mental-to-physical causation:

(3) *m* causes *p**.

*p** is a physical effect. Hence, (Closure) dictates that there is a physical cause of *p** at time *t*. Let *p* be such an event, which is an instantiation of *P* at time *t*.

(4) *p* causes *p**.

According to (Distinctness), *M* is not identical with *P*. Since events are instantiations *of* properties *by* objects *at* times, this means that *m* and *p* contain non-identical constituents. Hence,

(5) *m* is not identical with *p*.

Furthermore, according to (Strong supervenience), *M* has a physical property as its supervenience base. In this case, it is plausible to assume that *P* is this supervenience base. Hence,

(6) *m* has *p* as its supervenience base.

According to (3) and (4), *p** has two sufficient causes – *m* and *p* – occurring at the same time, *t*. However, this scenario cannot be categorized as genuine causal overdetermination. As we have seen, genuine casual overdetermination involves two or more *separate* and *independent* causal chains converging at a common effect. In this scenario, however, the causal chains leading from *m*-to-*p** and *p*-to-*p** are not independent because *m* supervenes on *p*. Moreover, genuine causal overdetermination requires that "each overdetermining cause plays a *distinct* and *distinctive* causal role" (Kim, 2005, 48) in producing the effect. However, it is not at all obvious what distinct and distinctive causal work *m* does in addition to that already done by *p*. Therefore, by (Exclusion), it follows that:

(7) Either *m* does not cause *p**, or *p* does not cause *p**.

But we cannot give up *p* as a cause of *p**, for *p*-to-*p** causation followed directly from (Closure). Therefore, *m*-to-*p** causation must be excluded. Hence,

(8) *m* does not cause *p**.

We have thus arrived at a contradiction between (3) and (8). This completes the second stage of our argument.

There are two lessons to be drawn from the exclusion argument. First, any mind-body theory that accepts all of the five premises is inconsistent, and, as a result, non-reductive physicalism must be rejected. Second, it is crucial to determine which premise(s) our best mind-body theory can do without. Here, Kim proposes that we retain all the other premises except (Distinctness). Do note that endorsing Kim's strategy amounts to embracing reductive physicalism. Non-reductive mind-body theories such as Cartesian substance dualism, emergentism, and parallelism, are all committed to (Distinctness). While some idealists may be willing to deny (Distinctness), idealism is incompatible with (Closure).

But why should we believe that rejecting (Distinctness) is the best solution to the exclusion argument? According to Kim, physicalists have only two choices when confronted with the exclusion argument – either reject (Distinctness) or reject (Causal Efficacy):

> The real aim of the [exclusion] argument, as far as my own philosophical interests are concerned, is not to show that mentality is epiphenomenal, or that mental causal relations are eliminated by physical causal relations; it is rather to show "either reduction or causal impotence." (...) If you deem yourself a physicalist (...) there are no other options. (Kim, 2005, 54-55)

We have already examined why epiphenomenalism is unsatisfactory in general. Causally impotent entities do not have any real interest or significance, to the extent that their removal would have no adverse effects. Furthermore, mental causation is particularly non-negotiable because it is crucial for the possibility of human agency, and hence for our moral practice. For voluntary actions to take place, mental states such as beliefs, desires, intentions, and decisions must cause bodily movements, thereby causing the rearrangement of objects around us. This consideration shows that reductive physicalism surpasses epiphenomenalism by a wide margin. For this reason, some philosophers, such as Papineau (2001, 8), consider the exclusion argument to be the strongest argument for reductive physicalism.[8]

---

[8] Before the exclusion argument became available, arguments for reductive physicalism were mostly based on simplicity considerations. In his classic paper, "Sensations and Brain Processes", Smart argues:

> Why do I wish to resist this suggestion [that a pain report, like "I am in pain", reports "an irreducibly physical something"]? Mainly because of Occam's razor. (...) That [sensations] should be correlated with brain processes does not help, for to say that they are correlated is to say that they are something "over and

However, Kim acknowledges that reductive physicalism is not a particularly appealing view of the mind either:

> If minds turn out to be mere configurations of neurons, silicon chips, or whatever and consciousness and thoughts are simply patterns of electrical activity in some groups of neurons, that doesn't seem much like saving minds as something distinctive, something we value, something that makes us the feeling, thinking, and rule-following creatures that we are. (Kim, 2005, 71)

Moreover, physicalists cannot completely avoid the threat of epiphenomenalism by simply "choosing" to reduce mental properties. This is because the exclusion argument does not establish the truth of reductive physicalism. As a result, depending on the model of reduction involved, a significant number of mental properties may ultimately turn out to be irreducible, rendering them epiphenomenal. This point is explicitly stated by Kim:

> The epiphenomenalist brunt of the [exclusion] argument is avoided if one is prepared, and is able, to choose the reductionist branch of the dilemma. It should be kept in mind that merely "choosing" reductionism doesn't make

---

> above." (...) That everything should be explicable in terms of physics (...) except the occurrences of sensations seems to me to be frankly unbelievable. Such sensations would be "nomological danglers", to use Feigl's expression. (Smart, 1959, 142)

However, it is dubious whether relying solely on Occam's razor has much persuasive force, as disputes regarding parsimony are prone to question-begging. What may appear as a redundancy or a "nomological danger" for reductive physicalists may be deemed essential for explaining the phenomena according to dualists.

> reductionism true; whether or not reductionism is sustainable as an option is an independent question that ought to be decided on its merits. (Kim, 2005, 55)

In fact, Kim ultimately concedes that phenomenal properties, or "qualia", such as experiences of pain and itch, are irreducible and thus epiphenomenal.[⑨] Given that Kim subscribes to Alexander's view that existence requires causal efficacy, this concession has the remarkable consequence that there are no such things as pains and itches.

Despite such shortcomings, Kim recommends that we endorse reductive physicalism because, according to the exclusion argument, it is the only plausible way of saving (most of) mental causation. Kim calls this position "conditional physical reductionism":

> The position we have arrived at may be called *conditional physical reductionism*: the thesis that if mental properties are to be causally efficacious, they must be physically reducible. That is, to save mental causation we must reduce mentality. (Kim, 2005, 5)

---

⑨ To be fair to Kim, he contends that the causal inefficacy of qualitative properties is not problematic because they are not required to save human agency. According to Kim, it is cognitive and intentional properties, such as believing and desiring, that form the foundation of human agency. And these properties *are* reducible, so human agency is saved under reductive physicalism after all. However, I am not entirely convinced that Kim's response, even if it is successful, resolves all concerns regarding qualia epiphenomenalism. For one thing, Kim (1984, 259; 2005, 9) subscribes to the causal theory of knowledge, according to which there must exist a causal relation between an object *o* and a subject *s* in order for *s* to gain knowledge of *o*. Therefore, qualia epiphenomenalism and the causal theory of knowledge collectively imply that we do not have any knowledge about qualia. But this consequence is absurd. We all know how pain feels! For a more detailed discussion of this matter, see Kim (2005), chapter 6.

In the next section, we will examine whether Kim succeeded in establishing conditional physical reductionism against non-reductive physicalists.

2.3. Arguments Against Conditional Physical Reductionism

Non-reductive physicalists were not pleased to hear that their position is incoherent and that they must choose between epiphenomenalism, reductionism, and anti-physicalism. As a result, they raised various objections to the exclusion argument. In this section, we will address two of these objections. First, the generalization objection argues that the exclusion argument threatens not only the causal efficacy of mental properties, but that of *all* special-science properties in general, such as chemical, biological, and geological properties. However, pervasive epiphenomenalism is unacceptable. Therefore, the objectors claim that there *must be* something wrong with the argument, although it is difficult to point out exactly where it goes wrong. Second, other non-reductive physicalists go an extra mile by attempting to deny one of the premises of the exclusion argument: (Exclusion). They do this by motivating the claim that an effect can have more than one sufficient cause if there is a tight modal connection between the causes. We will scrutinize each objection in turn and determine whether it succeeds in undermining conditional physical reductionism. By doing so, we will learn much about the nature and limitation of the exclusion argument.

2.3.1. The Generalization Objection

Numerous philosophers have pointed out that the premises of the exclusion argument have nothing to do with mentality *per se*. As a result, the argument can be generalized beyond mental properties to all special-science properties, including chemical, biological, and geological properties. To illustrate this point, let's consider biological

properties as an example. Many philosophers and biologists consider biology an *autonomous* science, "with its own distinctive methodology and system of concepts and not answerable to the methodological or explanatory constraints of fundamental [physics]" (Kim, 2010, 123).[10] Accordingly, biological properties are considered *non-physical* as they are not studied by fundamental physics, thereby satisfying the condition of (Distinctness). Furthermore, both biologists and philosophers acknowledge the causal efficacy of biological properties. Indeed, biology will hardly be a science worth pursuing if its objects of study are epiphenomena. Hence, biological properties meet the requirement of (Causal Efficacy). According to physicalism, any non-physical property must strongly supervene on physical properties. Therefore, biological properties also fulfill the criterion of (Strong Supervenience). Lastly, (Exclusion) and (Closure) are just general claims about the causal structure of the world. This consideration implies that biological causation is just as problematic as mental causation according to the exclusion argument. Since parallel considerations apply to other kinds of special-science properties, the exclusion argument leads to the conclusion that there is *nothing but* physical causation. Non-reductive physicalists take this consequence to be a *reductio* of the exclusion argument since non-physical causation, such as chemical and biological causation, evidently exist. In the exclusion argument literature, this strategy is called the "generalization objection", and it has been endorsed by various authors:

> Reserving causal status for strictly physical properties [in the manner of the exclusion argument] would make not only intentional properties epiphenomenal, it would also make the properties of chemistry, biology, neurophysiology and every theory outside of microphysics epiphenomenal. If the only sense in which intentional properties are epiphenomenal is a sense in which chemical and geological

[10] This view is forcefully advocated in Fodor's 1997 paper "Special Sciences: Still Autonomous After All These Years".

> properties are also epiphenomenal, need we have any real concern about their status; they seem to be in the best of company and no one seems worried about the causal status of chemical properties. (Van Gulick, 1992, 325)

> Moreover, I want to show that the metaphysical assumptions with which we began inevitably lead to scepticism not only about the efficacy of contentful thought, but about macro-causation generally. But if we lack warrant for claiming that macro-properties are generally causally relevant, and if we take explanations to mention causes, then most, if not all, of the putative explanations that are routinely offered and accepted in science and everyday life are not explanatory at all. (Baker, 1993, 77)

To recap, the reasoning behind the generalization objection is this: the exclusion argument renders mental properties epiphenomenal. But the exclusion argument applies not only to mental properties but to all non-physical special-science properties. Hence, all special-science properties are epiphenomenal. However, such pervasive epiphenomenalism is unacceptable. Therefore, the exclusion argument *must* be flawed.

However, proponents of the generalization objection commit a fatal mistake by claiming that the exclusion argument renders mental properties epiphenomenal. As we have observed, the exclusion argument does *not* suggest that mental properties are epiphenomenal. Rather, it forces physicalists to choose between mental epiphenomenalism and reduction. If other special-science properties are in the same boat as mental properties, then the same choice applies to them as well. Therefore, physicalists who find biological causation non-negotiable can always embrace the reduction of biological properties. To put the matter in another way, Kim can concede to the objectors that, according to the exclusion argument, there is *nothing but* physical causation. Consistently with this, he can argue that

biological causation exists because biological causation *just is* physical causation.

Some objectors may worry that reducing all special-science properties to retain their causal efficacy is problematic because it is too revisionist. Two replies can be made to this concern. First, it must be recognized that there is a strong initial intuition regarding the irreducibility of mental properties. Indeed, substance dualism reflects our folk conception of the mind better than any form of physicalism. However, it is unclear whether we have similarly robust dualist intuitions about biological or chemical properties. I do not think that many people would be shocked if it were discovered that the chemical property of [being a H_2O molecule] is reducible to basic physical properties. Second, even if the objectors are correct in arguing that reducing special-science properties goes against our intuitions, this does not constitute a valid objection to reductionism. Reduction, by its very nature, entails revising our ordinary or pre-theoretic conceptions of the world. Therefore, from the reductionist's perspective, this complaint will simply look question-begging. *If* all special-science properties can be systematically reduced, then no matter how this seems contrary to appearances, everyone must concede victory to reductive physicalism.

Now, we have examined that the exclusion argument does *not* establish the viability of reduction - merely "choosing" to reduce special-science properties does not mean that they thereby have been reduced. Hence, the generalization objection may have some bite if chemical or biological properties are particularly difficult to reduce. But this is unlikely given that mental properties are generally taken to be the biggest obstacle to physical reductionism. Thus, if reduction is untenable, this fact will already be manifest at the level of psychology, absolving us of the need to go further down to the level of biology and chemistry.

These considerations show that the generalization objection lacks persuasive force. Yet, it sheds light on the discussion of the exclusion argument by demonstrating that it is unnecessary to limit the debate solely to mental properties and mental causation. Instead, we can shift our attention to special-science properties in general and their causal efficacy. This is a significant improvement as it allows us to avoid introducing various idiosyncratic features of mentality such as first-person access, qualia, and intentionality into the debate, which would needlessly complicate matters. Therefore, in discussing the exclusion argument, I will refer to various cases of special-science causation as well as cases of mental causation. There is another lesson to be learnt from the generalization objection. The mind-body reductionism debate usually revolves around the reducibility of mental properties to neuropsychological properties. For example, reductive physicalists contend that the mental property of [being in pain] is reducible to the neurophysiological property of [C-fiber firing], whereas non-reductive physicalists deny this claim. However, even if this kind of reduction is possible, it does not automatically save mental causation. This is because neurophysiological properties themselves are subject to the exclusion argument. Therefore, to retain (Causal Efficacy), reductive physicalists must demonstrate how mental properties are reducible to *fundamental physical properties* whose causal efficacy is not threatened by the exclusion argument.

To conclude, the generalization objection advances our understanding of the exclusion argument by highlighting that we face a general metaphysical problem. However, without further argument, it fails to demonstrate that the exclusion argument consigns physicalism to pervasive epiphenomenalism.

2.3.2. Compatibilism: Denying (Exclusion)

In section 2.1, we noted that the truth of (Exclusion) is not intuitively evident:

> (Exclusion) No single event can have more than one sufficient cause occurring at any given time – unless it is a genuine case of causal overdetermination.

Indeed, many non-reductive physicalists have identified (Exclusion) as the source of the problem. Following Karen Bennett (2003; 2008), let us label such a non-reductionist position "causal compatibilism". So what does Kim have to say against compatibilism? Kim presents three arguments for (Exclusion), none of which are satisfactory.

First, in some places, Kim suggests that (Exclusion) is implicated in the meaning of "genuine causal overdetermination", making the principle "virtually an analytic truth with not much content" (Kim, 2005, 51). However, as Tim Crane and Steinvör Árnadóttir (2013, 257) pointed out, this is clearly mistaken. As we have examined, genuine causal overdetermination involves *independently* sufficient causes or independent causal chains converging on a single effect. Given such a characterization of genuine overdetermination, there is indeed a *weaker* analytic principle in the vicinity of (Exclusion):

> (Exclusion*) No single event can have more than one *independently* sufficient cause occurring at any given time – unless it is a genuine case of overdetermination.

However, all parties to the debate acknowledge that mental causation does *not* involve *independently* sufficient causes. Rather, the sufficient mental and physical causes are bound by a tight modal connection such as supervenience, which makes the mental cause *dependent on* the physical cause. Therefore, mental causation is not excluded by (Exclusion*). More importantly, in order to derive (Exclusion) from (Exclusion*), Kim must demonstrate that an event cannot have more than one sufficient cause *even if* there is a tight relation between the putative sufficient causes. This surely looks like a substantive claim in need of an argument. In fact, many compatibilists maintain that

(Exclusion) is false *because* an effect can have more than one sufficient cause if there is a tight modal connection between the causes.[11] To briefly introduce Bennett's brand of compatibilism, she argues that overdetermination requires the *non-vacuous* truth of certain counterfactuals. More precisely, in order for two causes m and p, to overdetermine some effect $p*$, it must be *non-vacuously* true that:

> (O1) if m had occurred without p, $p*$ would still have occurred: $(m \ \& \ \sim p) \ \Box\!\!\rightarrow \ p*$, and
>
> (O2) if p had occurred without m, $p*$ would still have occurred: $(p \ \& \ \sim m) \ \Box\!\!\rightarrow \ p*$.

Bennett asserts that the non-vacuous truth of these counterfactuals is necessary for genuine causal overdetermination because "they capture the reasoning we engage in when we want to distinguish cases of genuine overdetermination from cases of joint causation, or from cases in which one of the putative causes is not really a cause at all" (Bennett,

---

[11] Bennett (2003) claims that an effect is *not* overdetermined if there is a tight modal connection between the sufficient causes. Other causal compatibilists, such as Jonathan Schaffer (2003) and Theodore Sider (2003), concede that this is a case of overdetermination but argue that it is not problematic. However, as Bennett notes, the difference between these compatibilist positions is primarily a terminological issue, resulting from an equivocation on the term "overdetermination". Therefore, Bennett explicitly acknowledges that her compatibilism could be interpreted along the lines of Schaffer and Sider:

> The compatibilist could in principle accept that the effects of mental causes *are* always overdetermined, just not in a bad way – the overdetermination is perfectly acceptable, unsurprising, and unproblematic. This is just a terminological issue. For the sake of convenience, I shall speak as though the compatibilist wants to deny overdetermination altogether. (Bennett, 2003, 474)

2003, 477). Given this counterfactual test for overdetermination, it is easy to see that the modal connection between *m* and *p* is crucial in exempting mental causation from counting as an overdetermining cause. Since *m* supervenes on *p*, *p cannot* occur without *m*, rendering (O2) vacuously true. Now, whether Bennett's strategy succeeds in establishing compatibilism is a matter of ongoing debate. ⑫ Nevertheless, it must be acknowledged that Kim owes us an account of *why* Bennett's or other compatibilist tactics are unsatisfactory. As far as I know, Kim never addressed Bennett's strategy in detail. Judging from his works, however, I believe that he would argue that non-reductive physicalists cannot deny (Exclusion) because doing so leads to the violation of other premises, such as (Closure) or (Causal Efficacy). As we scrutinize Kim's second and third arguments for (Exclusion), we will evaluate these claims.

The second strategy employed by Kim to establish (Exclusion) was to argue that its rejection leads to the violation of (Closure). So let us assume that (Exclusion) is rejected, and hence that *m* and *p* are both sufficient causes of *p**. Here is a version of the argument which appeared in *Mind in a Physical World*:

> The [overdetermination] approach may come into conflict with the physical causal closure. For consider a world in which the physical cause [*p*] does not occur and which in other respects is as much like our world as possible. The overdetermination approach says that in such a world, the mental cause [*m*] causes a physical event [*p**]—namely that the principle of causal closure of the physical domain

⑫ A number of philosophers argue that Bennett's strategy is unsatisfactory because her test for overdetermination fail to deliver a necessary condition for causal overdetermination. For such claims, see Simona Aimar (2011), Chiwook Won (2014). From another angle, Sara Bernstein (2016, 37) objects that compatibilism merely "changes the subject from the contribution of the mental cause to the relationship between the physical and mental causes".

> no longer holds. I do not think we can accept this consequence: that a minimal counterfactual supposition like that can lead to a major change in the world. (Kim, 1998, 45)

This argument rests on two assumptions. The first assumption is that (Closure) should hold in non-actual worlds. I agree with Kim on this point. In fact, I believe that physicalism is committed to the idea that (Closure) is *at least* a nomologically necessary principle. The second assumption is that the counterfactual claim "if *p* had not occurred, m would still have occurred, thereby causing *p**" is true. This assumption, however, is false because it is not in line with our intuitions about counterfactuals. It is much more plausible to judge that if *p* had not occurred, *m* would not have occurred either. For instance, many will agree that "if the C-fiber in my brain had not been stimulated, I would not have felt any pain" is more plausible than "if the C-fiber in my brain had not been stimulated, I still would have felt pain, and this pain would have caused me to groan".

In his next book, *Physicalism, or Something Near Enough*, Kim concedes that his first argument was not quite right. Nevertheless, he makes the following response:

> In considering the claim that *m* and *p* are each a sufficient cause of *p**, however, we need to be able to consider a possible situation in which *m* occurs without *p* and evaluate the claim that in this possible situation *p** nonetheless follows. If such is not a possible situation—that is, if of necessity any non*p*-world is ipso facto a non*m*-world—what significance can we attach to the claim that *p* and *m* are each an overdetermining sufficient cause of *p**, that in addition to *p*, *m* also is a sufficient cause of *p**? (Kim, 2005, 46, the variables were changed for consistency)

The crux of Kim's reply is that if *m* is to count as a sufficient cause of *p**, *m* must be able to cause *p** in *p*'s absence. Therefore, let us assume that *W* is a world in which *m* occurs but *p* does not. It is plausible to think that *m*-to-*p** causation will remain intact in *W*. But does this constitute a violation of (Closure) in *W*, as Kim suggests?

Thomas Crisp and Ted Warfield (2001) provide an insightful discussion regarding this matter. First of all, we need to appreciate that (Strong Supervenience), which holds in the actual world, is a principle with a modal force:

> (Strong Supervenience) if any system *s* instantiates a mental property *M* at *t*, there *necessarily* exists a physical property *P* such that *s* instantiates *P* at *t*, and *necessarily* anything instantiating *P* at any time instantiates *M* at that time.

Now, for the purposes of the argument, let us assume, as Kim (2005, 49) does, that the kind of modality relevant to (Strong Supervenience) is nomological necessity.[13] Based on this assumption, we can infer that *W* must be either (i) a nomologically possible world where every mental event has a simultaneous physical event as its base, or (ii) a nomologically impossible world where mental events can occur autonomously. In case (i), *m* will have an alternative simultaneous physical event *p'* as its base, and *p'* will be causally sufficient for *p**. Therefore, (Closure) is not violated in *W*. On the other hand, in case (ii), *m* will not have any base physical event that will be causally

[13] Although Kim claims that "there are independent reasons for thinking that mind-body supervenience, if it holds, must be construed as nomological, not logical or metaphysical necessity" (Kim, 2005, 49), he does not state them. Without taking a side on this issue, I will simply note that this is a very controversial commitment. Loewer (2007, 244) and Bennett (2008, 286), for example, claim that mind-body supervenience requires metaphysical necessity.

sufficient for *p**. As a result, (Closure) is indeed violated in *W*. However, since *W* is a nomologically impossible world, it is unclear why physicalists should be concerned about the failure of (Closure) in such a remote world. Crisp and Warfield (2001, 314) conclude from these considerations that the rejection of (Exclusion) can indeed lead to the violation of (Closure), but this is not problematic because such violations only occur in possible worlds that are of no interest to physicalists.

At this point, Kim attempted to respond to each horn of Crisp and Warfield's dilemma. Regarding case (ii), Kim concedes that *W* is indeed a nomologically impossible world. Nevertheless, he objects that:

> *W* is nomologically impossible not because some physical law is violated in *W* but because some mental properties fail to supervene on physical properties—that is, because some psychophysical laws of our world fail in *W*. So *W* may well be a physically possible world; in fact, we may stipulate *W* to be a perfect duplicate of our world in all physical respects, including spacetime structure, basic physical laws, and fundamental particles. Should the physicalist not care whether physical causal closure holds in a world like *W*? Contrary to what Crisp and Warfield suggest, it seems obvious to me that anyone who cares about physicalism should care very much about (Closure) in *W*. (Kim, 2005, 49-50)

This response assumes that physicalists can, and should, distinguish between physically and nomologically possible worlds. Nomologically possible worlds are governed by the *same* natural laws as our world. In contrast, physically possible worlds only need to share all the basic physical laws that apply in our world. Therefore, they may differ from our world with respect to other natural laws, such as psychological or psychophysical laws. Now, if there are indeed nomologically impossible but physically possible worlds, and if *W* is such a world, as

Kim argues, then physicalists may have reason to worry about (Closure) in *W*.

However, I do not think that physicalists *can* draw such a distinction, for it leads to a decidedly anti-physicalist consequence that psychophysical laws are fundamental. For *reductio*, let us grant Kim that there is physical modality in addition to nomological and metaphysical modality. Since concepts of supervenience are distinguished by the notion of modality involved, we can introduce the corresponding idea of physical supervenience, in addition to nomological and metaphysical supervenience. Now, physicalists of all stripes concede that anything that fails to supervene on the physical must be recognized as a fundamental net addition to our ontology. Of course, physicalists believe that non-physical entities, if there are any, have only a derivative ontological status with respect to physical ones. This is why physicalists are committed to the idea that everything supervenes on the physical. Psychophysical laws are no exceptions to this rule.

So, *in what sense* can psychophysical laws be taken to supervene on the physical? Nomological supervenience immediately comes to mind. However, while it is true that psychophysical laws nomologically supervene on the physical, this is merely a *trivial* kind of supervenience because psychophysical laws themselves are included in the supervenience base. Indeed, psychophysical laws nomologically supervene on *anything whatsoever*. This means that nomological supervenience has no implication at all regarding the ontological status of psychophysical laws *vis-à-vis* the physical. Hence, physicalists must appeal to other kinds of supervenience to establish the ontological derivativeness of psychophysical laws. But Kim has already "set aside the possibility that mind-body supervenience is logically or metaphysically necessary" (Kim, 2005, 49). Furthermore, given Kim's admission that a perfect physical duplicate of our world can have different psychophysical laws from ours, psychophysical laws do not physically supervene on the physical. As there is no other

candidate notion of supervenience for physicalists to rely on, our consideration demonstrates that psychophysical laws turn out to be fundamental if we distinguish between physical and nomological modality. In my opinion, the only way Kim can prevent this untoward consequence is by conceding that psychophysical laws physically supervene on the physical.[14] However, this strategy implies that any physically possible world – a world whose basic physical laws are identical to those of our world – must have the same psychophysical laws as our world. Therefore, the purported distinction between physical and nomological modality collapses.

As regards case (i), Kim concedes that (Closure) is not violated in *W*. Yet, he replies that:

> In *W*, we have a replay of exactly the same situation with which we began—*m* has a physical base, *p*', threatening to preempt it as a cause of *p**. (...) As long as [(Strong Supervenience)] is held constant, there is no world in which *m* by itself, independently of a physical base, brings about *p**; whenever *m* claims to be a cause of *p**, there is some physical [event] waiting to claim at least an equal causal status. In the actual world, we may suppose that a continuous causal chain connects *p* with *p** (...). And it would be incoherent to suppose there is another causal chain from *m* to *p** that is independent of the causal process connecting *p* with *p**; the only plausible supposition is that

[14] Is it possible for Kim or any other physicalist to accept that psychophysical laws are fundamental? I do not see how. I believe that everyone will agree with Loewer's remark that "physicalism requires that once God created the totality of physical facts and laws he created the whole world. He didn't have to add mental (or any other) properties or *bridge laws* connecting them with physical properties or special-science laws connecting them with each other or extra causal relations or anything else" (Loewer, 2007, 244, my emphasis).

> if there is a causal path from *m* to *p**, that must coincide with the causal path from *p* to *p**. In *W*, another causal chain connects *p'* with *p**, and the *m*–*p** chain must coincide with that, and similarly in other such worlds. To be a cause of *p**, *m* must somehow ride piggyback on physical causal chains—distinct ones depending on which physical [event] subserves *m* on a given occasion, in the same world or in other possible worlds. And we may ask: In virtue of what relation it bears to physical [event] *p* does *m* earn its entitlement to a free ride on the causal chain from *p*-to-*p** and to claim this causal chain to be its own? Obviously, the only significant relation *m* bears to *p* is supervenience. But why should supervenience confer this right on *m*? The fact of the matter is that there is only one causal process here, from *p* to *p**, and *m*'s supposed causal contribution to the production of *p** is totally mysterious. In cases of [genuine] overdetermination, like two bullets hitting the victim's heart at the same time, the short circuit and the lantern causing a house fire, and so on, each overdetermining cause plays a distinct and distinctive causal role. The usual notion of overdetermination involves two or more separate and independent causal chains intersecting at a common effect. Because of (Strong Supervenience), however, this is not the kind of situation we have here. (...) Anyone tempted by the idea that mental events make their causal contributions by being overdetermining causes should reflect on whether this option could sufficiently vindicate the causal efficacy of the mental. (Kim, 2005, 47-49, the variables were changed for consistency)

I cited this lengthy passage because it is crucial to understanding the ultimate source of Kim's dissatisfaction with compatibilist tactics. He appears to believe that under compatibilism, mental causation becomes mysterious or too weak. This is an issue well worth pondering about. However, it is important to note that Kim's response to case (i),

whatever its own merits, simply changes the subject. This is because he explicitly acknowledges that (Closure) is not violated in case (i). Therefore, his response does nothing to defend his original argument that the rejection of (Exclusion) results in the violation of (Closure). For this reason, I will consider this response as Kim's third argument for (Exclusion).

In conclusion, Kim's second argument – rejection of (Exclusion) leads to the violation of (Closure) – is undermined because he failed to give a satisfactory response to Crisp and Warfield's dilemma: his response to case (i) changes the subject, while his response to case (ii) has implications that no physicalist can accept.

Let's now turn to Kim's third and last argument for (Exclusion), which asserts that its rejection leads to the violation of (Causal Efficacy). As we have seen in the passage just cited, Kim's primary concern with compatibilism is that it does not allow for the existence of an independent causal chain leading from *m*-to-*p**, leaving no room for the mental cause *m* to make any *independent* causal contribution to the production of the physical effect *p**. Rather, it appears that the physical cause *p*, which is *m*'s supervenience-base, is doing *all* the causal work, with *m* for some mysterious reason allowed to piggyback on the physical causal chain from *p*-to-*p**.[15] Let us say that a cause is

[15] Essentially the same objection against compatibilism is also made in Kim (1998):

> In cases of standard overdetermination, the overdetermining causes are *independent* events – two or more independent causal chains, each causally sufficient, converge upon a single effect. In contrast, in the case of [mental causation], we do not evidently have two independent causes: the instantiation of [the mental property] is dependent on the instantiation of [the physical property]. What isn't clear, however, is why this removes the difficulty: if the [physical event] is, in and of itself, a sufficient cause of the [effect], what *further* causal work is left for the

"independent" if it does independent causal work in bringing about the effect. It should be noted that independent causation is intimately tied to the *productive* or *generative* conception of causation:

> (The Productive/Generative Conception of Causation) A cause is something that produces, or generates, or brings about its effects, something from which the effects *derive* their existence or occurrence. (Kim, 2007, 235)

This conception of causation was given its classic expression when Elizabeth Anscombe (1993, 91) wrote, "causality consists in the derivativeness of an effect from its cause".[16]

Now, compatibilists readily concede that Kim is correct to point out that mental causes are not independent from the physical causes:

> I am happy to acknowledge that (...) the nonreductive physicalist does not have [any right to the] claim that the mental is *independently* causally efficacious. Perhaps doing without independent efficacy is a disturbing thought. But the fact is that it is a mistake to think that a physicalist can say

---

> [mental event]? The answer obviously is none: given the [physical event] as a full cause, there is no *additional* causal work left for [the mental event], or anything else. (...) The exclusion problem doesn't go away when we recognize the two purported causes as in some way related to each other, perhaps one being dependent on the other. As long as they are recognized as distinct events, each claiming to be a full cause of a single event, the problem remains. (Kim, 1998, 53, the example was changed for consistency)

[16] In his (2005, 18), Kim also states that the notion of causation he has "in mind is very close to the fundamental notion of causation, or determination, that I believe Elizabeth Anscombe was after in her *Causality and Determination*".

> anything else. (...) It is a direct consequence of their physicalism. (Bennett, 2008, 301)

In other words, compatibilists claim that mental causes are "dependent" in the sense that they are *dependently* causally efficacious. In cases of dependent causation, the cause does not do any "further causal work" (Kim, 1998, 53) in addition to that already performed by its dependence-base.

However, Kim objects that this notion of dependent causation is obscure, in contrast to the concept of independent causation. If an event *c* makes independent causal contribution to the production of another event *e*, then there is a clear sense in which *e* derives its existence from *c*. In addition, no mere epiphenomenon can play a distinct and distinctive causal role in bringing about *e*. Therefore, in cases of independent causation, the cause's status as a cause is unquestionable. On the other hand, a dependent cause's "supposed causal contribution to the production of its effect is totally mysterious" (Kim, 2005, 48). In fact, it does not seem to make any contribution at all. Therefore, in a case of dependent causation, there seems to be no sense in which the effect is derivative with respect to its cause. But then why should we regard dependent "causes" as causes at all? What distinguishes dependent causes from mere epiphenomena? Kim believes that compatibilists cannot provide any adequate response to these worries. That is why he asserts that dependent causation is "an empty verbal ploy; we can "say", if we want, that [*m*] is a "supervenient", "dependent", or "derivative" cause, or whatever (...). But this is only a gimmick with no meaning" (Kim, 2005, 62). Now, if Kim is correct to claim that there is no such thing as dependent causation, then compatibilism straightforwardly leads to mental epiphenomenalism according to the following argument. If compatibilism is true, then mental causes are dependent. However, there are no dependent causes. Therefore, there are no mental causes.

At this point, compatibilists can make several replies. Bennett argues that compatibilists do not have to vindicate dependent causation to respond to the exclusion argument. Instead, they just need to demonstrate that compatibilism is a viable option or that one can "choose" to deny (Exclusion):

> Responding to the exclusion problem requires less than is sometimes supposed. It does not require providing a positive story about how the mental manages to be causally efficacious. Telling such a story is of course required by a full defense of [dependent] mental causation from all challengers, but not by a defense from the exclusion problem in particular. (Bennett, 2008, 282)

At first glance, this may seem like a fair response. As we have seen, Kim argues that the exclusion argument compels us to "choose" reductive physicalism, but he acknowledges that the argument does not establish the truth of reductive physicalism. If no suitable model for the reduction of mental properties can be provided, then reductive physicalism must be abandoned. However, "whether or not reductionism is sustainable as an option is an independent question that ought to be decided on its merits" (Kim, 2005, 55). Bennet can claim that the same holds true for compatibilism. If the notion of dependent causation is indefensible, then compatibilism is undermined. But whether or not compatibilism is sustainable as an option is a separate issue.

Nonetheless, there are good reasons to believe that Bennett's response is nothing more than a Pyrrhic victory. If merely establishing the possibility of rejecting one of the premises constitutes a sufficient response to the exclusion argument, then no one should be bothered by the exclusion argument. For instance, epiphenomenalists can gladly "choose" to deny (Causal Efficacy), and Cartesian dualists can similarly "choose" to reject (Closure). However, nobody thinks that epiphenomenalists and Cartesian dualists can get off the hook *that*

easily. Physicalists contend that denying (Causal Efficacy) or (Closure) is not a genuine option because doing so would have extremely unpalatable consequences. Denying (Causal Efficacy) seems to endanger human agency, and many physicalists, such as Papineau (2001), argue that denying (Closure) conflicts with empirical evidence. This observation suggests that a complete response to the exclusion argument must include a defense of one's preferred option as a *plausible* resolution to the exclusion argument.[17] In this regard, it must be acknowledged that Kim was too hasty in concluding that reductive physicalism is the most plausible solution when compatibilism remains a genuine contender. Yet, it also must be recognized that compatibilists cannot simply show the consistency of denying (Exclusion) and walk

[17] My objection to Bennett gains much more traction when we consider that she regards the exclusion argument as an argument *for* physicalism:

> [Compatibilists] should not merely argue that we are not in trouble over the exclusion problem; we should argue that we are not in trouble while the dualist still is. (Bennett, 2008, 282)

But how is that possible, given that many dualists can simply reject (Closure)? She answers that:

> The question is not just whether dualists can *consistently* reject [(Closure)], but whether they can *plausibly* reject it. It is not clear that they need to endorse [(Closure)], but it is also not clear that they can happily deny it and walk away whistling. It is an interesting and important project, I think, to see whether even dualists have compelling reason to accept that physics is causally complete. (Bennett, 2008, 283)

However, if Bennett thinks that the exclusion argument can be used against dualists if they cannot *plausibly* reject (Closure), then why not apply the same standard to compatibilism? After all, Kim's objection is that (Exclusion) cannot be *plausibly* rejected because the coherence of dependent causation is strongly suspicious.

away whistling, without providing any answers to Kim's charge that their position leads to epiphenomenalism.

Therefore, I submit that compatibilists should take on Kim's challenge and seek to justify dependent causation.[18] As we have seen, Kim rejects dependent causation because it is incompatible with the productive conception of causation, which characterizes a cause as something that produces or generates its effects, with the effects *deriving* their existence or occurrence from it. Hence, to vindicate dependent causation, compatibilists must offer an alternative conception of causation that can accommodate it. Barry Loewer (2002; 2007) argues that the counterfactual conception of causation is suitable for this purpose. On this view, causal claims can be analyzed in terms of claims of counterfactual dependence. Counterfactual dependence is defined as follows: for actual events *c* and *e*, *e* counterfactually depends on *c* if and only if, if *c* had not occurred, *e* would not have occurred. Although I won't delve into the specifics of this account, it is easy to show how the counterfactual conception of causation can justify dependent causation. This is because *c* does not need to do any independent causal work in producing *e* for *e* to counterfactually depend on *c*. For instance, consider the following psychophysical counterfactual conditional claim (C) and its corresponding mind-body causal claim (D):

(C) If I had not had the desire to raise my hand, I would not have raised my hand.

(D) My desire to raise my hand caused my hand's rising.

[18] In their 2010 paper "Is Non-reductive Physicalism Viable within a Causal Powers Metaphysic?", Timothy O'Connor and Ross Churchill also argue that non-reductive physicalism is undermined by the exclusion argument if one assumes that productive account of causation.

Intuitively, (C) is true even if my desire to raise my hand played no independent causal role in bringing about my hand movement, but instead merely supervened on some neurophysiological event that did all the causal work. According to the counterfactual account of causation, dependent mental causation is thereby established because the truth of the causal claim (D) is underwritten by the truth of (C).

However, Kim (2007, 234) considers this consequence to be a defect rather than a virtue of the counterfactual approach to causation because even epiphenomenalists who deny the truth of (D) can nonetheless acknowledge the truth of (C). In other words, Kim objects that the existence of a mere counterfactual dependence between *c* and *e* cannot guarantee the existence of a causal relation between *c* and *e*. In response, Loewer (2007, 257) argues that the counterfactual account of causation demands a very particular way of evaluating counterfactuals, such as David Lewis's (1973) semantics for counterfactuals. On this ground, he claims that (D) will come out false under epiphenomenalism if we evaluate counterfactuals along Lewisian lines. Those interested in the details of the debate between Loewer and Kim can refer to the cited papers. Here, I note only that neither Kim (2007) nor Loewer (2007) provides a conclusive justification for their preferred account of causation.[19] Moreover, it is

[19] Kim and Loewer themselves concede this point:

> I don't expect Kim to be persuaded by my counterfactual defense of causation because he considers it, at best, causation lite as compared to causation as production. (Loewer, 2002, 661)

> We care about mental causation because we care about *human agency*, and agency requires the productive/generative conception of causation. I don't have a knockdown argument to prove that agency requires productive causation; I hope what I will say here makes my claim at least plausible. (Kim, 2007, 236)

unlikely that a satisfying account of the metaphysics of causation will appear in the foreseeable future. Therefore, the dispute between Kim and non-reductive physicalists over the truth of (Exclusion) will not be resolved soon.

I have argued at length that Kim's arguments for (Exclusion) are at best inconclusive. The first argument, which claims that (Exclusion) is an analytic principle, and the second argument, which claims that the rejection of (Exclusion) leads to the violation of (Closure), were decisively refuted.[20] In contrast, the third argument, which states that the rejection of (Exclusion) results in the denial of (Causal Efficacy), is quite forceful. However, this strategy requires a commitment to a particular metaphysics of causation, namely the "robust, “thick” concept of productive or generative causation rather than a “thin” concept based on the idea of counterfactual dependence or simple Humean “constant conjunctions"" (Kim, 2005, 38).[21] Since it remains

---

[20] To be more charitable to Kim regarding the first argument, it can be acknowledged that (Exclusion) is indeed an analytic truth within the framework of the productive conception of causation. This is because dependent causation does not make sense under this conception. Therefore, if we interpret Kim's first argument in this manner, it should be considered as a highly misleading formulation of the third argument.

[21] As a side note, Kim (2005; 2007) suggests that the productive conception of causation is friendly to, and perhaps even accounted for by, Phil Dowe (2000) and Wesley Salmon (1994)'s conserved quantity approach to causation:

> I am implicitly asking the reader to think causation in terms of actual productive/generative mechanisms involving energy flow, momentum transfer, and the like, and not merely in terms of counterfactual dependencies. (Kim, 2005, 47)

> [Productive causation] involves a *real connectedness* between cause and effect, and the connection is constituted by phenomena such as energy flow and momentum transfer, an

to be seen whether the productive conception of causation is correct, the third argument remains inconclusive. This highlights a limitation of the exclusion argument, as it is powerless to prevent those non-reductive physicalists who are willing to endorse dependent causation from rejecting (Exclusion).

At this point, there is another issue worth considering. Suppose that one prefers the productive conception of causation and thereby agrees with Kim's verdict that dependent causation is only a gimmick with no meaning. *In this case,* should she follow Kim in "choosing" reductive physicalism? In other words, is denying (Distinctness) the best solution to the exclusion argument for those who are congenial to (Exclusion)? I will give a negative answer to this question in chapter 3.

---

actual movement of some (conserved) physical quantity (Kim, 2007, 236).

However, the mere movement of causal "oomph" or some conserved physical quantity from *c* to *e* does *not* necessarily entail that *c* produces *e*, and that *e* derives existence from *c*. This point becomes clear when we consider that compatibilism can accommodate the conserved quantity approach to causation. "The trick would be to claim that mental property instances (or events, etc.) and their physical realizers *only provide one injection of oomph*. (...) To see the idea, imagine two events, one a proper part of the other, such that the part constitutes what might be called an 'efficacious core': the other parts of the larger event are wholly inert. One might well want to say that both the larger and the smaller event are causally sufficient for some effect, but do not overdetermine it" (Bennett, 2008, 294). Loewer (2007, 258) makes a similar point. Therefore, advocates of the productive conception of causation should argue that even the conserved quantity approach to causation is too "thin" and not robust enough to ground productive causal relations.

Chapter 3. An Argument Against Weak Conditional Physical Reductionism: Given (Exclusion), Denying (Distinctness) Entails the Denial of (Causal Efficacy)

> (Distinctness) Mental properties are distinct from or are not identical with, physical properties.
>
> (Causal Efficacy) Mental properties have causal efficacy – that is, instantiations of a mental property *M* can, and do, cause other properties to be instantiated *in virtue of being an instance of M*.

In section 2.3.3, we learned that the exclusion argument is not as powerful as initially thought. This is because (Exclusion) is not a metaphysically neutral principle, requiring a commitment to the productive conception of causation. This consideration shows that Kim failed to establish conditional physical reductionism using the exclusion argument:

> What we have established, if our considerations have been generally correct, is a *conditional thesis*, "If mentality is to have a causal influence in the physical domain—in fact, if it is to have any causal efficacy at all—it must be physically reducible." I have not argued for reductionism simpliciter; rather, I have argued that mental causation requires reduction, and that anyone who believes in mental causation must be prepared to endorse mind-body reduction. We may call this "conditional [physical] reductionism." (Kim, 2005, 161, my emphasis).

The *conditional thesis* is false because mentality can have causal influence in the physical domain *even if* it is physically irreducible, *as long as* the coherence of dependent causation can be established.

Even so, there is a *weaker conditional thesis* in the offing, which states, "If mentality is to have a *productive* causal influence in the physical domain—in fact, if it is to have any causal efficacy at all—it must be physically reducible." It is an interesting question whether Kim, using the exclusion argument, can establish this weaker kind of conditional physical reductionism. The issue could be given a slightly different formulation. Given (Exclusion), is denying (Distinctness) the best, or even perhaps the only plausible option for retaining (Causal Efficacy)?

In this chapter, I will argue that Kim cannot even defend *weak conditional physical reductionism* against anti-physicalism. One possible way of doing this is by providing an external critique of reductive physicalism. That is, by arguing that denying (Closure) is more plausible than denying (Distinctness). My strategy, however, has to do with the internal consequences of reductive physicalism. An important point to note regarding the exclusion argument is that denying one of the premises does not necessarily mean that the other premises are preserved. Indeed, Kim's second and third argument for (Exclusion) were that its denial leads to the violation of another premise, such as (Closure) or (Causal Efficacy). The same applies to the denial of (Distinctness). One cannot simply assume that she has secured (Causal Efficacy) by denying (Distinctness). Thus, my argument against *weak* conditional physical reductionism is the following: Given (Exclusion) or the productive conception of causation, denying (Distinctness) leads to the denial of (Causal Efficacy).

In other words, under the productive conception of causation, reductive physicalism is not an alternative to, but rather a version of epiphenomenalism. This is not an argument against reductionism *simpliciter* because I do *not* claim that (Distinctness) cannot be denied. Rather, I argue that mental or other special-science properties cannot be reduced in a way that preserves their causal efficacy, according to the productive conception of causation. If my argument is on the right

track, it suggests that anyone who accepts the productive conception of causation should look outside physicalism to save mental causation.

Furthermore, I have no objection against those reductive physicalists who do not subscribe to the productive conception but instead to the nomological or counterfactual conceptions of causation. Nevertheless, I submit that such reductive physicalists cannot use the exclusion argument to motivate their position. They will have to go back to Herbert Feigl (1958) and Smart (1959)'s arguments for reductive physicalism based on simplicity considerations. However, the persuasive force of Occam's razor is dubious, to say the least.

3.1. The Premises

To proceed with my argument, let us start by taking the denial of (Distinctness) as our first premise. Suppose that M is an arbitrary mental property. Then,

(9) A mental property M is identical with a physical property.

Next, I will introduce the well-known distinction between abundant and sparse conceptions of properties ("property" is here understood to cover both properties and relations). "Any class of things, be it ever so gerrymandered and miscellaneous and indescribable in thought and language, and be it ever so superfluous in characterising the world, is nevertheless a[n] [abundant] property" (Lewis, 1983, 346). Hence, negative properties such as *not being golden*, disjunctive properties such as *being green or blue*, and extremely gerrymandered properties such as *quadding*, which is just like addition unless one of its operands is 57 or greater, in which case it always yields 5, all belong to the category of abundant properties. Lewis claims that we need abundant properties to play the role of semantic values of meaningful predicates. However, such an abundant rabble of properties is unsuited to perform many other tasks expected of properties:

> Because [abundant] properties are so abundant, they are undiscriminating. Any two things share infinitely many [abundant] properties, and fail to share infinitely many others. That is so whether the two things are perfect duplicates or utterly dissimilar. Thus [abundant] properties do nothing to capture facts of resemblance. That is work more suited to the sparse universals. Likewise, [abundant] properties do nothing to capture the causal powers of things. Almost all [abundant] properties are causally irrelevant, and there is nothing to make the relevant ones stand out from the crowd. [Abundant] properties carve reality at the joints -- and everywhere else as well. If it's distinctions we want, too much structure is no better than none. It would be otherwise if we had not only the countless throng of all properties, but also an elite minority of special properties. (Lewis, 1983, 346)

Lewis calls these elite minorities of special properties "sparse properties". To recapitulate, sparse properties serve three crucial metaphysical functions:

> (i) *Similarity*: sparse properties ground objective similarities.
>
> (ii) *Causality*: sparse properties track or carve out causal powers.
>
> (iii) *Minimality*: sparse properties characterize things completely and without redundancy.

Of relevance to our discussion is feature (ii). According to Lewis, as stated in the aforementioned passage, only sparse properties have causal powers – merely abundant properties are causally irrelevant or inefficacious. From this we get the second premise of our argument:

(10) If *M* is not a sparse property, then *M* is causally inefficacious.

(9) and (10) collectively imply that reductive physicalism cannot save mental causation unless each mental property can be identified with some sparse physical property. Admittedly, this is quite a challenging requirement. Because of this, I believe that Kim's attitude toward (10) lacks consistency. In some of his writings, he strongly advocates it.[22] For example, in the following passage, Kim explicitly supports the distinction between abundant and sparse conceptions of properties. Moreover, he concedes to (10)'s requirement that mental properties must be sparse if there is to be mental causation:

> I am advocating here what is called a "sparse" conception of properties as distinguished from the "latitudinarian" or "abundant" conception.[23] (...) I believe it is clear, although I will not belabor the point, that the conception of properties appropriate to the present context [regarding the causal efficacy of mental properties] is the sparse one. In fact current debates over the mind-body problem and mental causation tacitly presuppose a particularly robust version

[22] Apart from the cited passage below, see also Kim (1992, 24-25).

[23] Kim sometimes endorses the stronger position that only sparse properties deserve to be considered "properties" in a proper sense, whereas abundant properties are better classified as "concepts" or "descriptions". For example, after showing that second-order functional "properties" do not possess the necessary causal unity required of sparse properties, he claims that "it is less misleading to speak of second-order *descriptions* or *designators* of properties, or second-order *concepts*, than second-order properties" (Kim, 1998, 104). However, in discussing (10), we need not be concerned with Kim's anti-realism about abundant properties. What matters for the purposes of my argument is that he accepts the *distinction* between abundant and sparse conceptions of properties, regardless of their ontological standing.

> of this approach according to which differences in [sparse] properties must reflect differences in causal powers. (Kim, 1998, 105)

However, as we will see in section 3.3.3., Kim asserts that functional properties are causally efficacious despite not being sparse. And I will defend (10) against this claim in due course. But for now, let us take (10) for granted.

Now, we must figure out which kinds of physical properties are sparse. But what are physical properties in the first place? The truth of physicalism – whether everything is physical or not – is arguably the most central issue in the philosophy of mind. Ironically, providing a definition of the term 'physical' that can establish physicalism as a philosophically significant doctrine turned out to be exceedingly difficult. Crane and David Hugh Mellor (1990) present the difficulty in the form of a dilemma. If we define the term 'physical' via reference to present-day physics, then physicalism is clearly false. This is because it is highly probable that contemporary physics is false, and that future physics will identify new objects and properties. On the other hand, if we choose to define the term 'physical' with respect to the *completed* ideal physics of some unspecified future, then physicalism is true, but only trivially so. Who can predict what items such an ideal physics will contain? For all we know, this completed future science may well include souls and mental forces as its objects of study. If Crane and Mellor are correct, then there is no question of physicalism since it is either clearly false or trivially true. *A fortiori*, there is no question of *reductive* physicalism.

Of course, I do not intend to claim here that rejecting (Distinctness) is a non-starter because reductive physicalism is an obscure position.

On the contrary, I would like to adopt Kim's solution to Crane and Mellor's dilemma:[24]

> Let us begin by considering the idea of a 'physical property'. I am not here seeking a definition or a general criterion. The question is rather this: Assuming that the properties and magnitudes that figure in basic physics are physical properties, what other properties are to be counted as members of the physical domain? When we speak of the physical, or physical properties, in discussing the mind-body problem, we standardly include chemical, biological, and neural properties among physical properties. Without invoking a general definition of 'physical', can we give some principled ground for this practice? And when we speak of the causal closure of the physical domain, just what should be included in the physical domain, and why? We assume that the entities and properties of basic physics are in this domain, but what else goes in there and why? (Kim, 1997, 293)

Kim acknowledges that it is extremely difficult to provide a general definition of 'physical' and hence does not attempt to provide one. Yet, he claims that we have a fairly firm grasp of the concept. This is shown by the fact that we call some things 'physical' and other things not. I agree with Kim on this point. The fact that we cannot give an explicit definition of a concept does not mean that we do not have *any* understanding of the concept at all. There are many concepts (e.g., morally wrong, blameworthy) we understand but do not know how to analyze. Nevertheless, we can provide an informative account of the

[24] To avoid misunderstandings, Kim (1997, 293) does not directly address Crane and Mellor's dilemma. Nevertheless, Kim's claim in the cited page does seem to provide a plausible answer to the problem. If you think that Kim himself would have preferred a different solution, read it as *my* solution to Crane and Mellor's dilemma, inspired by Kim.

meaning of a term by giving a principled explanation of the way we use it. So how does Kim propose to account for our practice of using the term "physical"? Here is the basic strategy: let us call the objects and properties studied by contemporary physics "basic physical entities". Basic physical entities comprise objects like quarks and leptons and properties like mass, charge, and spin. Basic physical entities are indisputably included in the physical domain. Now, we also include other objects and properties in the physical domain *because* they are certain kinds of complexes or combinations *built up* from the basic physical entities. According to Kim, there are three such ways of building up the physical domain (Kim calls such ways "closure conditions"):

> These, then, are three closure conditions: first, any entity aggregated out of physical entities is physical; second, any property that is micro-based on entities and properties in the physical domain is also physical; third, any property defined as a second-order property over physical properties is physical. Are there other closure conditions? I am not sure. Conjunctive properties can be taken as a special case of micro-based properties (if we can waive the condition that the constituents of such properties must be nonoverlapping proper constituents): having P & Q is being composed of parts a_1 and a_2, where $a_1 = a_2$, such that a_1 is P and a_2 is Q. But disjunctions and complementations are not in yet; these operations give rise to some well-known complications that need not be discussed here. (Kim, 1997, 294)

A lot of technical terms appear here, and I will give a detailed explanation of each condition in a moment. But let us first extract from the cited passage the following premise:

> (11) Any physical property is a (i) basic physical property, or (ii) a property that is micro-based on entities and

properties in the physical domain[25], or (iii) a second-order property over physical properties.

(11) is false if there are yet undiscovered closure conditions of the physical domain. But Kim will agree with me that this is unlikely.

Now, basic physical properties are the paradigms of sparse properties.[26] However, it is evident that whatever physical property *M* is reducible to, it is not a basic physical property. This is because the basic objects and properties studied by contemporary physics (e.g., leptons, quarks, spin, charge) belong to the microscopic realm, whereas *M* is a macro-property, possessed only by macroscopic objects like animals. This means that if *M* can be identified with any physical property at all, it must be a highly complex macro-property built up from these basic physical properties. Hence, we get:

(12) *M* is not a basic physical property.

We now need to find out whether micro-based properties or second-order properties are sparse. I will carry out this task by taking a closer look at each of Kim's three closure conditions of the physical domain. The first condition – any entity aggregated out of physical entities is physical – is quite straightforward. Some physicalists claim that the physical domain only comprises microphysical entities and their properties. Against this excessively narrow conception, Kim argues that "the physical domain must also include aggregates of basic particles, aggregates of these aggregates, and so on, without end; atoms, molecules, cells, tables, organisms, mountains, planets, and all the rest belong, without question, in the physical domain" (Kim, 1997,

[25] Following Kim (1997, 294), I will include conjunctive properties in this category.

[26] Hence Lewis (1986, 356-357) says, "physics is relevant because it aspires to give an inventory of natural [or sparse] properties".

293). Since we are here dealing with properties, not objects, the first condition need not occupy us further.

The second condition – any property that is micro-based on entities and properties in the physical domain is also physical – makes use of David Armstrong's (1997, 31-37) notion of micro-structural or micro-based property. Kim proposes the following characterization of micro-based properties:

> P is a *micro-based property* just in case P is the property of having proper parts, a_1, a_2, …, a_n, such that $P_1(a_1)$, $P_2(a_2)$, …, $P_n(a_n)$, and $R(a_1, \ldots, a_n)$. (Kim, 1997, 292)

Kim (1997, 292; 2005, 57) argues that many properties possessed by macro-objects which satisfy the first closure condition (e.g., atoms, molecules, cells, etc.) are micro-based. For instance, the property of [being a H_2O molecule] is a micro-based property because it is the property of [having two hydrogen atoms and one oxygen in a such-and-such bonding relation]. The property of [having a mass of 10 kg] is another example of a micro-based property because it is the property of [being made up of proper parts, a_i, each with a mass of m_i, where the m_is sum to 10 kg]. There are two features of micro-based properties worth mentioning. First, micro-based properties are macro-properties in the sense that they are instantiated by the whole object and not by the whole's (micro) proper parts a_1, a_2, …, a_n. For example, [being a H_2O molecule] is instantiated by a whole molecule, not by the hydrogen or oxygen atoms that constitute it. Second, a micro-based property endows *new causal powers* to its bearer that go beyond the causal powers of its micro-constituents.[27] Consider again

[27] In section 3.3.2, it will be demonstrated that this claim is not strictly true. *Physical* micro-based properties do bring new causal powers, as they are conjunctions of basic physical entities. However, other micro-based properties, such as [being a H_2O molecule], are merely vast disjunctions of physical micro-based properties. And it will be demonstrated in section

the micro-based property of [being a H_2O molecule]. A H_2O molecule which instantiates this property has the power to extinguish flames. But none of its proper parts possesses this power. They rather have powers to the contrary: a hydrogen atom is flammable, and an oxygen atom helps other things burn. Similar things can be said as regards [having a mass of 10 kg]. Suppose that there is a pressure plate that is activated only when a weight of 10kg or more is applied. In this case, a bowling ball weighing 10kg has the power to activate the plate, while none of its parts possesses this ability. If micro-based properties have new causal powers, then they are undoubtedly causally efficacious and are required to track causal powers. Consequently, micro-based properties are sparse. This seems to make micro-based properties ideal reduction-bases for mental properties.

But the question is: can mental properties be identified with properties that are micro-based on entities and properties in the physical domain? My answer is no. First, note that the micro-based properties which serve as the reduction-base for mental properties must be *micro-based on entities and properties in the physical domain*. Let us call such micro-based properties "physical micro-based properties". The next thing to note is that the fact that a property *P* is a micro-based property does *not* imply the fact that *P* is a *physical* micro-based property. In other words, some (in fact, most) micro-based properties are *non-physical*. Cartesian substance dualism provides a striking illustration of this point. According to Descartes, a person is a union of a mind and a body. Although it is unclear how Descartes envisioned the union of the mind and body[28], for the sake of simplicity, let us assume that this union is achieved by a primitive mind-body union relation *R*. Then, under the Cartesian ontology, the property of [being a person] is a micro-based property, as it is the

---

3.2.2. that disjunctive properties do not bring any new causal powers into the world.

[28] For Kim's interpretation of Descartes' mind-body theory, see his (2005), chapter 3.

property of [having a mind and a body in a primitive union relation *R*]. However, [being a person] is evidently not a *physical* micro-based property because it is (partially) built up from non-physical entities such as the mind and the mind-body union relation *R*.

Then what distinguishes physical micro-based properties from non-physical ones? Answer: physical micro-based properties are *conjunctions* of basic physical entities.[29] This point becomes obvious when we consider that micro-basing is a way of building up the physical domain by using the basic physical entities as foundational building blocks. Consequently, all of the proper parts a_1, a_2, …, a_n and the properties P_1, P_2, ..., P_n which constitute a physical micro-based property must be basic physical entities. In addition, it is required that R(a_1, a_2, …, a_n) be composed of basic physical relations – ideally pairwise relations between basic physical objects. This means that the instances of a physical micro-based property will share a highly specific micro-configuration. In fact, Kim concedes to this point:

> [Physical micro-based properties] supervene on *specific mereological configurations* involving these microproperties—for a rather obvious and uninteresting reason: they *are* identical with these micro-configurations.[30] (Kim, 1998, 117-118)

What Kim didn't realize, however, is that this feature prevents the identification of *any* macro-property in general, as commonly understood or studied by the special sciences, with a physical micro-

[29] This is why, as we have seen, Kim classifies conjunctive properties as a species of micro-based properties.

[30] To be precise, Kim does not differentiate between physical micro-based properties and micro-based properties in general. However, as Kim is discussing micro-based properties that fall within the physical realm, I have taken the liberty of inserting the term "physical micro-based property" in the cited passage.

based property. This is because instances of most macro-properties do *not* share a specific micro-configuration. On the contrary, such macro-properties must be understood as vast disjunctions of physical micro-based properties.[3132] In effect, I am here arguing against Kim that quotidian and scientific macro-properties such as [having a mass of 10 kg] and [being a H_2O molecule] fail to qualify as physical micro-based properties. Take the property of [having a mass of 10kg] as an example. The micro-configurations of objects (e.g., bowling balls, drawers, 2-year-old children, desks, etc.) that weigh 10kg vary to a bewildering degree. As for the property of [being a H_2O molecule], Kim identifies it with the property of [having two hydrogen atoms and one oxygen in a such-and-such bonding relation]. Now, this is an accurate description and shows that [being a H_2O molecule] is a micro-based property. However, it does not demonstrate that [being a H_2O molecule] is a *physical* micro-based property since the properties [being a hydrogen atom], [being an oxygen atom], and [being in a bonding relation] are themselves vast disjunctions of physical micro-based properties. Accordingly, the micro-based property of [having two hydrogen atoms and one oxygen in a such-and-such bonding relation] that is composed of these properties is also a vast disjunction of physical micro-based properties. Now, to show that [being a H_2O molecule] is a physical micro-based property,

---

[31] This point was made by various authors. Ned Block (2003, 145-146) argues that the macro-property of [being jade] cannot be identified with any (physical) micro-based property because it can be micro-based in both the property of [being nephrite] and the property of [being jadeite]. Schaffer (2004, 96) likewise argues that "*the conjunctive/structural model is the wrong way to understand the macro-scientific properties*. Consider the property of being a desire. It is not a conjunction of, or structure of, fundamental properties – it is a disjunction of such conjunctions/structures."

[32] Here, one might suggest modifying Kim's closure condition to include disjunctions of physical micro-based properties in the physical domain. My reply to this objection is that disjunctive properties are not sparse and, therefore, lack causal efficacy regardless of whether they are considered physical or not. This idea is developed in more detail in section 3.3.2.

it must be identified with the property of having some specific quantum configuration. However, as Loewer noted, this is not possible:

> Being a water molecule is not an aggregate or conjunction of fundamental microphysical properties but a vast disjunction since water molecules can occupy infinitely many quantum states. (Loewer, 2002, 656)

If my preceding comments about micro-based properties are correct, then identifying mental properties with physical micro-based properties is an untenable strategy. Type-identity theorists typically aim to identify mental properties such as [being in pain] with neurophysiological properties, such as [C-fiber firing]. If such neurophysiological reductions are achievable, they would demonstrate that [being in pain] is a micro-based property. However, this does *not* imply that [being in pain] is a *physical* micro-based property, as the property of [C-fiber firing] itself is not a *physical* micro-based property but a vast disjunction thereof. To prove that [being in pain] is a *physical* micro-based property, it must be shown that it is identical to the property of [having Q], where Q denotes an extremely specific micro-configuration involving basic physical entities. However, even the most hardcore type-identity theorist will acknowledge that such a reduction is impossible. Hence, we get:

> (13) M is not a property that is micro-based on entities and properties in the physical domain. In other words, M is not a physical micro-based property.

The final closure condition – any property defined as a second-order property over physical properties is physical – pertains to second-order properties over physical properties. Let D be a set of first-order physical properties. Then a second-order property over physical properties can be defined using existential quantification as "the property of having some property R in D satisfying a certain condition C." Second-order properties are important because

functionalists typically construe functional properties as second-order properties over physical properties, where the relevant condition *C* is causal. Here, the physical properties that satisfy the causal condition *C* are said to "realize" the functional property. Suppose, for instance, that the property of [being a gene] is a functional property. Then it could be defined as "the property of having some property that performs a certain causal function, namely that of transmitting phenotypic characteristics from parents to offsprings. As it turns out, it is the DNA molecule that fills this causal specification" (Kim, 1999, 10). In this case, the property of [being a DNA molecule] is said to realize the property of [being a gene].

Now, many functionalists are non-reductive physicalists. Consequently, they will reject the idea of regarding functional properties as physical *simply because* they are defined by existential quantification over physical properties. Instead, they will argue that functional properties, due to their multiple realizability, cannot be reduced to their physical realizers and thus constitute an autonomous domain for the special sciences. Of course, Kim does not think that multiple realizability is an impediment to the reduction of functional properties. In effect, Kim proposes various ways of reducing second-order properties to *other* physical properties and we will scrutinize them in section 3.3.1. For now, without getting into the details of this dispute, I will adopt Kim's stance of counting any second-order property over physical properties as physical. If non-reductive physicalists are correct about the irreducibility of functional properties, then (11) could simply be substituted with (11'):

> (11') Any physical property is a (i) basic physical property, or (ii) a property that is micro-based on entities and properties in the physical domain.

And by employing (11') instead of (11) as a premise, it is easier to demonstrate my point that mental properties cannot be reduced to any causally efficacious physical property. This is because we have

already established that mental properties can neither be identified with a basic physical property nor a physical micro-based property. Nevertheless, working with (11) grants Kim a dialectical advantage and therefore allows me to draw a more robust conclusion.

Many philosophers have equated mental properties with functional properties. Accordingly, if mental properties can be identified with a physical property at all, functional properties appear to be the most suitable candidates. The problem, however, is that functional properties are not sparse. This is because functional properties merely "pick out" physical properties that fulfill a certain causal specification. Hence, they do not introduce new causal powers into the world. In other words, if an object possesses a realizer property *R* of a functional property *F*, it does not acquire additional powers *in virtue of* having *F* beyond those already provided by *R*. I believe that Kim agrees with me on this point:

> By existential quantification over a given domain of properties, we do not literally bring into being a new set of [sparse] properties. That would be sheer magic, especially if we adopt the plausible view that distinct [sparse] properties must represent distinct causal powers. (Kim, 1998, 103)

Now, opponents of the productive conception of causation may argue that a property *does not* need to introduce new causal powers to be sparse. To elaborate, these objectors will concede to Kim that distinct sparse properties must represent distinct causal powers. However, they will maintain that a second-order property represents *distinct* causal powers with regard to its realizer properties *not because* it represents certain causal powers not possessed by its realizers, but rather *because* it represents a *proper subset* of the causal powers of its realizers. Sydney Shoemaker puts the point in the following way:

> Property *X* realizes property *Y* just in case the (...) powers bestowed by *Y* are a subset of the (...) powers bestowed by *X* (...). Where the realized property is multiply realizable, the (...) powers bestowed by it will be a proper subset of the sets bestowed by each of the realizer properties.[33] (Shoemaker, 2001, 78-79)

Of course, according to this strategy, the causal efficacy of second-order properties is *redundant* because all of their causal powers are already possessed by their realizers. However, Jonathan Schaffer claims that causal redundancy should not dissuade philosophers from regarding second-order properties as sparse:

> Why can't nature contain redundancies (...)? Surely it is metaphysically possible that nature itself could be nonminimal.[34] And a redundant world could still enjoy objective similarities and causal powers, and could still be assayed. (Schaffer, 2004, 99)

For the purpose of my argument, it is not necessary to discuss whether this strategy successfully makes second-order properties sparse.[35] It only needs to be acknowledged that this strategy requires a commitment to redundant or dependent causation. That is, in whatever sense a second-order property can be said to possess a proper subset of the causal powers of its realizers, it is not the power to *produce* anything. Otherwise, each case of second-order property

---

[33] A similar strategy was also defended in Wilson (2015).

[34] Here Schaffer is referring to infinitely complex worlds where properties are endlessly supervenient upon lower-level properties.

[35] Kim, for example, does not agree with Schaffer that the metaphysical possibility of infinitely complex nonminimal worlds vindicates redundant/dependent causation. According to Kim, causation in infinitely complex worlds can be preserved as long as the series of identity claims goes on indefinitely, such that: "$M_L = M_{L-1} = M_{L-2} = M_{L-3}$ ..." (Kim, 2005, 69).

causation would be a case of genuine causal overdetermination where there are two causal chains, one from the realizer property-instance and another from the second-order property-instance, converging on the same effect. But no physicalist envisions the situation this way. All will agree that "the fact of the matter is that there is only one causal process here, from" (Kim, 2005, 48) the realizer property-instance to the effect, with the second-order "cause" somehow riding piggyback on this causal chain. Therefore, anyone who shares Schaffer's intuition regarding redundant causal powers should renounce their commitment to the productive conception of causation. Equivalently, reductive physicalists who endorse the productive conception of causation have no choice but to accept the following premise:

> (14) If *M* is a second-order property over physical properties, then *M* is not a sparse property.

3.2. The Argument

We have gathered all the premises necessary to establish that, under the productive conception of causation, reductive physicalism is incapable of preserving mental causation, as mental properties cannot be identified with any causally efficacious physical property. The structure of the argument to this effect is straightforward. Reductive physicalism is committed to (9):

> (9) A mental property *M* is identical with a physical property. [Premise]

(10) follows from the widely accepted distinction between abundant and sparse conceptions of properties:

> (10) If *M* is not a sparse property, then *M* is causally inefficacious. [Premise]

The domain of physical properties which serve as the reduction base of mental properties is built up from the basic physical properties in the following way:

> (11) Any physical property is a (i) basic physical property, or (ii) a property that is micro-based on entities and properties in the physical domain (= a physical micro-based property), or (iii) a second-order property over physical properties. [Premise]

However, for reasons already stated:

> (12) *M* is not a basic physical property. [Premise]

> (13) *M* is not a property that is micro-based on entities and properties in the physical domain. In other words, *M* is not a physical micro-based property. [Premise]

Hence,

> (15) *M* is a second-order property over physical properties.[36] [From (9), (11), (12), and (13)]

However, under the productive conception of causation:

> (14) If *M* is a second-order property over physical properties, then *M* is not a sparse property. [Premise]

---

[36] I believe that Kim will agree with (15). Kim (1998; 1999; 2005) has long advocated the functional model of reduction, according to which the reduction of a property *P* consists in *functionalizing P*. Here, to functionalize *P* is to (re)construe *P* as a functional property. And he claims that the "functionalization of a property is both necessary and sufficient for reduction" (Kim, 1999, 18).

Hence,

(16) *M* is not a sparse property. [From (14) and (15)]

Therefore,

(17) *M* is causally inefficacious. [From (10) and (16)]

Since *M* is an arbitrary mental property, (17) implies the violation of (Causal Efficacy). Now, in section 2.3.1, we saw that the premises of the exclusion argument have nothing do with mentality *per se*. This implied that all special-science properties are susceptible to the exclusion argument, and that reductive physicalists must reduce all of them to retain their causal efficacy. My argument is also entirely general in nature, as none of its premises relied on the idiosyncratic features of mentality. Therefore, while I presented the argument with the assumption that *M* is a mental property, this is not a requirement. It is easy to see that by replacing *M* with biological or chemical properties, such as [being a H_2O molecule] and [C-fiber firing], the same result can be achieved. Consequently, under the productive conception of causation, reductive physicalism indeed leads to pervasive epiphenomenalism.

3.3. Objections and Replies

Naturally, Kim and other reductive physicalists will not let me have my way so easily. Significant and insightful objections can be raised against (9), (10), and (11). Therefore, I will address each of these concerns in turn and defend my premises against them.

3.3.1. Denying (9): Conservative Reduction and Eliminative Reduction

Kim may complain that I am working under an unduly narrow conception of reduction. To elaborate, reduction can take either a conservative or an eliminative form. Conservative approaches to reduction retain the existence of the reduced entities. Identification is a paradigmatic instance of conservative reduction because if X is reduced to Y through identification ($X = Y$), the existence of X is conserved. In contrast, eliminative approaches to reduction remove the reduced entities from our ontology. Hence, eliminative reduction has no need for identities. Nonetheless, both are legitimate approaches to reduction because they result in a leaner ontology.

As an illustration, consider second-order properties. In presenting my argument, I conceded to Kim that any second-order property over physical properties can be regarded as a physical property. As we have observed, however, this is a contentious position. Non-reductive physicalists will refuse to include second-order properties over physical properties into the physical domain unless it can be demonstrated that they are reducible to *other* indisputably physical properties. In response, Kim contends that second-order properties are amenable to both conservative and eliminative reduction.[37] Let us assume that M is a second-order property. Then, a conservative reduction of M involves identifying it with the disjunction of its first-order physical realizer properties, (R_1 v R_2 v ...). It's worth noting that this identity is metaphysically contingent but nomologically necessary. This is because M is defined in terms of a causal condition, and whether a property satisfies such conditions depends on the laws that hold at a given world. At this point, one may wonder why Kim claims that second-order properties can be physically reduced by identifying them with disjunctive properties when he excluded disjunctive properties from the physical domain. However, it's worth clarifying that Kim is only discussing how second-order properties *could* be conservatively reduced, if possible. In fact, he leans towards

[37] Kim endorses this strategy in various works. See, for example, Kim (1999, 15-18; 2005, 58).

the eliminative approach when it comes to second-order properties.[38] Moreover, Kim's position on disjunctive properties is quite nuanced. We will discuss this matter in the next section. For now, however, let's set the issue aside and turn to the eliminative reduction of second-order properties. Eliminative reduction of a second-order property *M* involves rejecting *M* as a genuine property and recognizing only the expression "*M*" or the concept *M*. According to this approach, existentially quantifying over first-order properties does not create any new property; it simply produces new ways of selecting or grouping first-order properties based on causal specifications that are of epistemic or practical interest.[39]

Now, the existence of two approaches to reduction suggests that there are two methods of refuting (Distinctness):

> (Distinctness) Mental properties are distinct from or are not identical with, physical properties.

First, using the conservative approach, we can deny (Distinctness) by identifying mental properties with physical properties. Second, using the eliminative approach, we can deny (Distinctness) by getting rid of mental properties. This strategy is viable because *X* cannot be distinct from *Y* if *X* does not exist at all. However, (9) is inadequate as it only takes the conservative strategy into account:

---

[38] For example, Kim claims that we should eschew "the talk of functional *properties* in favor of functional *concepts* and *expressions*" (Kim, 1998, 110).

[39] Endorsing the eliminative strategy means embracing anti-realism when it comes to abundant properties. This is evident when we realize that any set of things constitutes an abundant property. Therefore, if one is a realist about abundant properties, then there would be no reason to reject the existence of second-order properties.

(9) A mental property *M* is identical with a physical property *P*.

Therefore, according to the objection, (9) should be substituted by (9*):

(9*) A mental property *M* identical with a physical property *P*, or *M* does not exist.

In response to this objection, I acknowledge that eliminative reduction is an adequate form of reduction. However, I fail to see how invoking the eliminative strategy can aid the reductive physicalist in preserving the causal efficacy of mental properties. It is evident that if we eliminate an entity *X* from our ontology, we also eliminate *X*'s putative causal powers. Of course, there will be concepts and expressions in the wake of eliminative reductions. However, unlike properties, concepts or expressions are causally irrelevant. In fact, Kim himself seems to accept this point. In the following passage, Kim argues that the micro-based property of [being jade] is amenable to either conservative or eliminative reduction.[40] Of the two methods, Kim favors the conservative approach *because* it preserves the causal efficacy of [being jade]. This implies that [being jade] will become causally inert or irrelevant if it is reduced through elimination:

> We can either deny that jade is a genuine kind (at least, jade is not a kind of mineral), on account of its causal heterogeneity, or identify jade with a disjunctive kind, jadeite or nephrite (that is, being jade is identified with having the microstructure of jadeite or the microstructure of nephrite). The second option which allows disjunctive kinds is a more conservative approach and may be more viable as a general solution. On the disjunctive approach, being jade turns out to be a causally heterogeneous property, not *a causally inert one*, and jade turns out to be

[40] Note that [being jade] is *not* a *physical* micro-based property.

a causally heterogeneous kind, not *a causally irrelevant one*. (Kim, 2005, 58, my emphasis)

I therefore submit that Kim *must* appeal to conservative reduction to uphold mental causation. At this point, Kim might suggest a number of eclectic approaches that combine some form of conservative reduction of mental entities with the eliminative reduction of mental properties. First, there is the local reduction strategy.[41] Suppose that M is a functional *concept* that is multiply realized by properties R_1, R_2, ..., R_n. Then there are species-specific or structure-specific contexts C_1, C_2, ..., C_n such that in each context C_i, only one realizer property P_i realizes M. Accordingly, if we countenance properties such as [M in C_1], [M in C_2], ..., [M in C_n], then a weaker form of type-identity theory can still be maintained by endorsing property identity statements like [M in C_1] = R_1, [M in C_2] = R_2, ..., [M in C_n] = R_n. For instance, let us suppose that <being in pain> is a functional concept that is realized by the neurophysiological property of [C-fiber firing] in humans and by [O-fiber firing] in octopuses. In this case, while there is no property of [being in pain], there are properties like [pain-in-humans], [pain-in-octopuses] such that [pain-in-humans] is [C-fiber excitation] and [pain-in-octopuses] is [O-fiber excitation]. "In this way multiply realized properties are sundered into their diverse realizers in different species and structures" (Kim, 1998, 111).

However, the local reduction strategy is suspiciously *ad hoc* since there is no reason to believe in the existence of localized properties such as [M in C_i]. To illustrate this point, let us first recall that any physical property is a (i) basic physical property, or (ii) a property that is micro-based on entities and properties in the physical domain (= a physical micro-based property), or (iii) a second-order property over physical properties. Therefore, the first-order physical properties P_i's that realize a functional concept M are either basic

[41] In Kim (1992, 19-26), he articulates and defends the local reduction strategy in detail.

physical properties or physical micro-based properties. Since *M* is a macro-concept, *M* will have physical micro-based properties as its realizers. We must now consider just *how specific* the structural context C_i must be to identify the localized mental property [*M* in C_i] with M's micro-based realizer P_i. In our example, it was suggested that species-specific contexts, such as "in-humans" or "in-octopuses", could provide us with localized mental properties, such as [pain-in-humans] or [pain-in-octopuses]. Now, if such a species-specific localization is successful, then there is something to be said for the local reduction strategy. True, eliminating general mental properties such as [being in pain] implies forfeiting psychology as a special science studying mental properties that are shared across species. Nevertheless, much generality is retained if we have properties like [pain-in-humans], and *human* psychology can still be a special science that studies such localized mental properties. Unfortunately, species-specific contexts are too crude for local reductions. Even at the neurophysiological level, [pain-in-humans] is multiply realized due to neuroplasticity. It is possible that a patient who suffers irrecoverable damage in the C-fiber area retains her capacity to feel pain, as this function may be played by some other part of the brain as a result of neuron pathways being rewired. Furthermore, even if we grant that [C-fiber firing] is the unique *neural* realizer of [pain-in-human], identifying the two properties does *not* lead to a *physical* reduction of [pain-in-human]. This is because [C-fiber firing] is a vast disjunction of physical micro-based properties that realize *M*. As it turns out, the property of [pain-in-human] must be eliminated just like the property of [being in pain]. Hence, the localization process must be iterated until we reach the highly specific micro-based properties P_i's that realize *M*. At this point, it becomes clear that the relevant structure-specific context C_i could not be anything other than P_i. In other words, the structural context C_i in which P_i uniquely realizes *M just is* being in P_i. If C_i is in any way more general, the correlated localized property [*M* in C_i] will be multiply realized by the physical micro-based properties P_i's and hence will be subject to elimination. Consequently, the "properties" which result from the local

reduction of M are [M in P_1], [M in P_2], ... In this scenario, M has been disintegrated to such an extent that it seems unmotivated and even absurd to consider [M in P_i] a mental property. While it makes sense to talk of *human* psychology, it is absurd to talk of P_i -psychology, where P_i is a property of having some highly specific micro-configuration. Moreover, why should we countenance such things as [M in P_i] in the first place? Obviously, [M in P_i] has no place in ordinary contexts. Nor is it useful in scientific contexts, as it completely lacks generality. In my opinion, if anything is a concept or an expression, then it is <M in P_i>, or "M in P_i", as it merely picks out a causal feature of P_i that is of interest to us.[42]

Let us now move onto the second eclectic approach, the token identity strategy. Token identity theories conservatively reduce mental property instances to physical property instances by way of identification. Perhaps it could be maintained that this is consistent with the elimination of mental properties. Although it is not entirely clear, Kim could be interpreted as endorsing this strategy in certain places. In the following paragraph, Kim seems to suggest that conservatively reducing the instances of [being jade] is reduction enough:

> Each instance of jade—that is, each individual piece of jade—is either jadeite or nephrite, and I don't see anything wrong about identifying its being jade with its being nephrite (if it is nephrite) or with its being jadeite (if it's jadeite). (...) *All we need is identity at the level of instances, not necessarily at the level of kinds and properties*; causation after all is a relation between property or kind-instances, not between properties or kinds as such. (Kim, 2005, 58, my emphasis)

[42] A similar objection against the local reduction strategy was raised by Ausonio Marras (2002, 248).

I have two objections against the token-identity strategy. First, the coherence of the strategy can be questioned as it is not clear how there could be instances of a *property M* when there is no *M* at all. Of course, there will be instances of a *concept M* or an *expression* "*M*". However, these items cannot save *M*-causation because instances of concepts or expressions do not enter into causal relations. Therefore, much more needs to be said to make this strategy work. Second, this strategy eliminates mental properties. Hence, regardless of what it says about the status of mental property-instances, mental properties are rendered causally irrelevant. In other words, this strategy does nothing to save (Causal Efficacy):

> (Causal Efficacy) Mental properties have causal efficacy – that is, instantiations of a mental property *M* can, and do, cause other properties to be instantiated *in virtue of being an instance of M*.

As a result, if Kim or other reductive physicalists wish to pursue this strategy, they must argue that mental causation can be saved *even if* (Causal Efficacy) is false. Perhaps this is what Kim had in mind when he said that "causation after all is a relation between property or kind-instances, not between properties or kinds as such". However, this approach would be vulnerable to the same criticism that undermined anomalous monism: it leads to type-epiphenomenalism, even though it is not a form of token-epiphenomenalism. Given that Kim was one of the most vocal critics of anomalous monism, he would not welcome this consequence. Indeed, in the following passage, Kim makes it clear that reductive physicalism can, and should, save the causal efficacy of mental properties:

> The position we have arrived at may be called *conditional physical reductionism*: the thesis that if *mental properties* are to be causally efficacious, they must be physically reducible. That is, to save mental causation we must reduce mentality. (Kim, 2005, 5, my emphasis)

3.3.2. Denying (11): Letting Disjunctive Properties into the Physical Domain

Our discussion in section 3.3.1 demonstrated that appealing to an alternative method of reduction, namely, eliminative reduction, cannot save reductive physicalists from the threat of epiphenomenalism. They must adhere to conservative reduction, as stated in (9). Another approach that reductive physicalists may consider is expanding the reduction-base. As we have seen, Kim (1997) proposed three closure conditions of the physical domain, which was captured in (11):

> (11) Any physical property is a (i) basic physical property, or (ii) a property that is micro-based on entities and properties in the physical domain (= a physical micro-based property), or (iii) a second-order property over physical properties.

In this formulation, disjunctions of physical properties are excluded from the physical domain. However, some may argue that reductive physicalists shouldn't shun disjunctive properties for the following reasons. Firstly, excluding disjunctive properties from the physical domain renders most macro-properties as *non-physical*, since macro-properties that appear in ordinary and in scientific contexts, such as [being a H_2O molecule] or [C-fiber firing], are vast disjunctions of *physical* micro-based properties. This is problematic because, as Kim himself acknowledges, "when we speak of the physical, or physical properties, (...) we standardly include chemical, biological, and neural properties among physical properties" (Kim, 1997, 293). Secondly, non-reductive physicalists refuse to acknowledge (iii) as a closure condition of the physical domain *unless* there are independent reasons to believe that second-order properties can be physically reduced. In response to such a requirement, reductive physicalists can demonstrate the conservative reducibility of a functional property *only*

by identifying it with the disjunction of its physical realizer properties. However, if *no* disjunctive property is physical, as (11) claims, then the path to the conservative reduction of functional properties is blocked, and reductive physicalists have no other option but to eliminate them. However, as we have previously observed, eliminating a property renders it causally inefficacious. Therefore, reductive physicalists should strive to secure the causal efficacy of second-order properties by conservatively reducing them to disjunctive properties. Given these considerations, one might suggest that we modify Kim's closure condition and replace (11) with (11*):

> (11*) Any physical property is a (i) basic physical property, or (ii) a property that is micro-based on entities and properties in the physical domain (= a physical micro-based property), or (iii) a second-order property over physical properties, or (iv) a disjunctive property with all of its disjuncts being properties in the physical domain.

It is not within my interest to determine which of (11) and (11*) better reflects our intuitive conception of the physical domain. Therefore, I will assume that the objector is correct in accepting (11*). However, for this modification to be useful in responding to my argument, the objector must establish the truth of (18):

> (18) *M* is a disjunctive property with all of its disjuncts being properties in the physical domain, and *M* is sparse.

In my opinion, there are two ways in which reductive physicalists might attempt to do this. First, they can identify *M* with a functional property and then reduce it to the disjunction of its physical realizers. This move will be endorsed by functionalists with reductive inclinations. Second, they can identify *M* with disjunctions of *physical* micro-based properties, such as [C-fiber firing]. This move will be endorsed by type-identity theorists. However, I will argue that neither of these strategies succeeds, *not because* such reductions are

impossible, but rather because neither disjunctions of physical realizers nor disjunctions of physical micro-based properties are sparse properties.

Let us first deal with disjunctions of physical realizers. Thankfully, Kim (1992, 11-13; 1998, 106-112) himself presented a simple but powerful argument demonstrating that the disjunctive properties that could serve as the reduction-base of functional properties are *not* sparse. Suppose that M is a functional property and R_1, R_2, ... are its first-order physical realizer properties. Then,

(19) $M = (R_1 \text{ v } R_2 \text{ v } ...)$. [Premise]

(20) $(R_1 \text{ v } R_2 \text{ v } ...)$ is a disjunction of *heterogeneous* properties. [Premise]

(21) A sparse property is a nomically projectible property. [Premise]

(22) A disjunction of heterogeneous properties is not a nomically projectible property. [Premise]

(23) Therefore, $(R_1 \text{ v } R_2 \text{ v } ...)$ is not a sparse property. [From (19), (20), (21), and (22)]

The argument is valid. So let's examine each premise in turn. Premise (19) is true because it simply states that a functional property M is conservatively reduced by identifying it with the disjunction of its first-order physical realizer properties, $(R_1 \text{ v } R_2 \text{ v } ...)$. The truth of premise (20) will be evident if we consider the physical realizers of, say, [being a heart]. Any organ that pumps blood qualifies as a heart. And the physical characteristics of human hearts can differ greatly from those of hearts in birds or reptiles. Furthermore, some hearts are

artificial and don't consist of flesh at all.[43] (21) appeals to the notion of a "nomically projectible property". X is projectible if X is confirmable by observations of positive instances. And laws are distinguished from mere empirical generalizations because they are projectible. Therefore, by a "nomically projectible property", Kim is referring to a property that can figure in laws. Now, the truth of premise (21) is apparent when we acknowledge that sparse properties are causally efficacious properties, and that laws underwrite causal relations. In other words, a sparse property is nomically projectible because a property incapable of figuring in laws is not causally efficacious. The most crucial premise requiring support is premise (22). Kim provides an argument in its favor, which proceeds as follows:

Suppose that the following law (L1) holds:

(L1) Patients with arthritis have painful joints.

If disjunctions of heterogenous properties are nomically projectible, then they can figure in laws. Hence, we can choose an arbitrary disease (say, lupus) and formulate the following putative disjunctive law (DL):

(DL) Patients with either arthritis or lupus have painful joints.

Anyone who has arthritis also has either arthritis or lupus. Thus, if the disjunctive property of [having either arthritis or lupus] is nomically

[43] To be sure, some non-reductive physicalists like Block (1997) argue against premise (b). He claims that, in some cases, the physical realizers of a functional property may not be so heterogeneous because laws of nature may put significant constraints on what kinds of physical structure can play the functional role. I will not attempt to defend Kim regarding this matter because I present a stronger argument that demonstrates that *no* disjunctive property is sparse.

projectible, then each datum that confirms (L1) also confirms (DL). However, (DL) is logically equivalent to the conjunction of (L1) and (L2)[44]:

(L2) Patients with lupus have painful joints.

As a consequence, we have the absurd result that the data confirming the truth of statement (L1) would also confirm the truth of statement (L2). Who would have imagined that by simply examining persons with arthritis, one could make surprising discoveries about persons with lupus! The argument shows that (DL) is not a law, which in turn implies the truth of premise (d): heterogeneous disjunctions such as [having either arthritis or lupus] are not nomically projectible.

Kim's argument shows that *even if* functional properties can be conservatively reduced to the disjunctions of their realizers, they are not sparse. I concur with Kim on this point. For those who are congenial to the productive conception of causation, regardless of how functional properties are conceived – whether as existential quantifications, concepts, or disjunctions – they are causally inefficacious.

Let us now turn our attention to the disjunctions of physical micro-based properties. In this case, we cannot utilize Kim's argument from (19) to (23) to establish their non-sparseness. In Kim's argument, the *heterogeneity* of the disjunction plays a crucial role. However, many disjunctions of physical micro-based properties, such as [being a H_2O molecule] or [C-fiber firing], are relatively *homogeneous* properties because their instances have a high degree of structural similarity.[45] For example, while the *physical* micro-based disjuncts that make up

[44] Formally, $\Box(((P \vee Q) \rightarrow R) \leftrightarrow ((P \rightarrow R) \& (Q \rightarrow R)))$.

[45] Of course, the degree of structural similarity of instances of [being a H_2O molecule] or [C-fiber firing] is *not* high enough to classify them as *physical* micro-based properties.

the disjunctive property of [being a H_2O molecule] are heterogeneous when viewed from a microphysical perspective, given that H_2O molecules can occupy a vast number of quantum states, they are nonetheless homogeneous from a chemical standpoint. To be specific, each *physical* micro-based disjunct involves two hydrogen atoms and an oxygen atom being arranged in a particular bonding relation. This distinguishes [being a H_2O molecule] from other miscellaneous and arbitrary kinds of disjunctive properties like [having either arthritis or lupus], where the disjuncts are heterogenous from *every* perspective.

Armed with this distinction, one could claim that the disjunction of *homogenous* properties is sparse *because* it is nomically projectible. For instance, (LD') is a genuine law:

(LD') H_2O molecules put out flames.[46]

Now, suppose that C_1 and C_2 are two of the physical micro-based disjuncts that constitute [being a H_2O molecule]. Then, (L3) and (L4) are laws:

(L3) Objects that have C_1 put out flames.

(L4) Objects that have C_2 put out flames.

Given that (LD') logically implies the truth of both (L3) and (L4), it follows that any empirical evidence confirming the truth of (L3) would also confirm the truth of (L4). This result, however, is neither absurd nor surprising. In fact, it is entirely reasonable to deduce the causal potential of one particular micro-configuration that constitutes H_2O from another such configuration. This strategy gains additional support from the observation that properties like [being a H_2O molecule]

[46] (LD') is not precisely accurate since it is not a single H_2O molecule, but rather a vast quantity of H_2O molecules that can extinguish flames. However, this approximation is sufficient for explanatory purposes.

appear to confer novel causal powers. For example, H_2O molecules have the power to put out flames, whereas hydrogen and oxygen atoms do not.

In fact, this may have been Kim's view regarding disjunctive properties. In his (1997), from which we obtained (11), Kim refused to let *any* disjunction of physical properties into the physical domain. In his other works, however, his stance on disjunctive properties is less extreme:

> There is nothing wrong with disjunctive predicates as such; the trouble arises when the kinds denoted by the disjoined predicates are *heterogeneous*, "wildly disjunctive", so that instances falling under them do not show the kind of "similarity", or unity, that we expect of instances falling under a single kind. (...) Disjunctive properties, unlike conjunctive properties, do not guarantee similarity for instances falling under them. And similarity, it is said, is the core of our idea of a [sparse] property. (...) The point about disjunctive properties is best put as a closure condition on [sparse] properties: the class of [sparse] properties is not closed under disjunction (presumably, nor under negation). Thus, there may well be [sparse] properties P and Q such that P or Q is also a [sparse] property, but its being so doesn't follow from the mere fact that P and Q are properties. (Kim, 1992, 13, my emphasis)

Here, Kim seems to indicate that disjunctive properties are sparse so long as their disjuncts are homogeneous, ensuring nomic projectibility and a relatively high degree of similarity among their instances.

These considerations show that I need to present another argument to demonstrate the non-sparseness of *homogeneous* disjunctive properties. Before doing so, however, I want to point out that reductive physicalists cannot argue that a property is sparse simply *because* it

is nomically projectible. To be sure, (21), its converse, is true: if a property is sparse, then it is nomically projectible. However, not all properties capable of figuring in laws are sparse. To claim otherwise would be to endorse the nomological conception of causation, according to which causation can be derived out of lawful regularities. However, in the context of our argument, reductive physicalists are working under the assumption that the productive conception of causation is true, which is antithetical to the nomological view. Indeed, in his (2007), siding with the productive conception of causation, Kim makes the following objection to Fodor's defense of special-science causation based on the nomological conception of causation:

> Though there may be projectible special-science properties and there may be special-science laws, that does not guarantee that there is causation in the special sciences. (...) To be sure, if there are *causal* laws in psychology, they will license ascription of causal responsibility to psychological properties and ground psychological causal relations. The crucial question unaddressed by Fodor is whether psychological laws *are* causal laws – that is, whether the regularities we observe in the psychological domain are causal regularities, or mere reflections of the causal regularities at a more fundamental level. (Kim, 2007, 232)

This consideration shows that homogeneous disjunctive properties like [being a H_2O molecule] can be non-sparse, despite being nomically projectible.

Now, I will present an argument which demonstrates that *no* disjunctive property, whether homogeneous or not, is sparse. The idea is that the same line of reasoning involved in the exclusion argument can be used to exclude the putative causal efficacy of *any* disjunctive

property.[47] Let (P v Q) be an arbitrary disjunction of physical micro-based properties, where P and Q may be homogeneous.[48] Now, if (P v Q) is causally efficacious, its instantiations must cause other properties to be instantiated *in virtue of* being an instance of (P v Q). Let E be such a property. Then,

(24) (P v Q) causes E.

For simplicity, in presenting the argument, I will say things like "(P v Q) causes E". But this is short for "an instance of (P v Q) causes an instance of E". Now, it is evident that,

(25) (P v Q) supervenes on P.

Assuming that (P v Q) was instantiated in this case *because* P was instantiated, it follows that P causes E. Otherwise, we would have to accept the idea that disjunctions can introduce a new set of causal powers, which would be tantamount to believing in magic. For example, an H_2O molecule h possesses the power to extinguish flames. Could we say that the property of [Being an H_2O molecule] bestowed this power on h *independently* of h's specific mereological configuration, in virtue of which h counts as a H_2O molecule? Clearly, the answer is no. This implies that P has whatever causal powers (P v Q) has. Therefore, we get,

(26) P causes E.[49]

---

[47] A similar objection was made by Sven Walter (2008, 686).

[48] (P v Q) was chosen for the sake of simplicity. My argument remains effective no matter how complex the disjunction is, as long as its disjuncts are physical micro-based properties.

[49] It should be noted that the structure of my argument is not *exactly* the same as that of the exclusion argument. One cannot rely on (Closure) to derive the truth of (26). This is because we are operating on the assumption that disjunctions of physical properties, such as (P v Q), are themselves physical properties.

Then, (Exclusion) straightforwardly applies to this case. According to (24) and (25), *E* has two sufficient causes – (*P* v *Q*) and *P* – occurring at the same time. But this scenario is not a case of genuine causal overdetermination because it does not involve two *separate* and *independent* causal chains converging at a common effect. This is because (*P* v *Q*) supervenes on *P.* Therefore, by (Exclusion), it follows that,

(27) Either (*P* v *Q*) does not cause *E* or *P* does not cause *E*.

Is it possible to exclude *P*-to-*E* causation in favor of (*P* v *Q*)-to-*E* causation? No. Recall that *P* is a physical micro-based property. And physical micro-based properties are sparse because they are conjunctions of basic physical properties. Moreover, in discussing (25), it was demonstrated that *P* must have whatever causal powers (*P* v *Q*) has. Consequently, we cannot choose to make *P* causally inefficacious. Therefore,

(28) (*P* v *Q*) does not cause *E*.

Even if we do not appeal to any explicit argument such as this, anyone friendly to the productive conception of causation should have strong suspicions about the causal efficacy of disjunctions of physical micro-based properties. According to this conception, a cause must play a distinct and distinctive role in the production of the effect. But what *independent* causal work can a disjunction of physical micro-based properties do in addition to that already done by its disjunct? The fact of the matter is that there is only one causal process here, from *P*-to-*E*, and (*P* v *Q*)'s supposed causal contribution to the production of *E* is totally mysterious. Indeed, one can make sense of "disjunctive causation" only by defending the notion of dependent causation, which is a "mere gimmick with no meaning" (Kim, 2005, 62) according to the proponents of the productive conception of causation.

Now, if my argument is correct, it undermines the supposed causal efficacy of *all* disjunctive properties, including [being a H_2O molecule] and [C-fiber firing]. Some reductive physicalists might object that this consequence is too skeptical to take seriously. After all, properties like [being a H_2O molecule] do appear to confer novel causal powers on their bearer. For example, H_2O molecules have the power to put out flames, whereas hydrogen and oxygen atoms do not. However, this claim rests on a confusion. I have no objection to the idea that an H_2O molecule *h* has more causal powers than its parts. But this does *not* imply that the property of [being a H_2O molecule] is the *source* of these additional causal powers. Rather, *h* has more causal powers than its parts *only because h* instantiates a specific mereological configuration or a physical micro-based property which is a disjunct of [being a H_2O molecule].

However, even if my response is adequate, the objectors may still remain unsatisfied. They may insist that the causal efficacy of special-science properties, such as [being a H_2O molecule] and [C-fiber firing], is non-negotiable. To such reductive physicalists, I suggest that they reconsider either their commitment to physicalism or to the productive conception of causation. Physicalism holds that any property not studied by contemporary physics is derivative and hence "built up" from basic physical properties. Under such a view, there is no choice but to consider properties like [being a H_2O molecule] and [C-fiber firing] as disjunctive properties, which undermines their causal efficacy under the productive conception of causation. However, if physicalism is abandoned, then one may seek to retain the causal efficacy of special-science properties by regarding them as fundamental. If this option seems "kooky", then they should attempt to vindicate the causal efficacy of disjunctive properties by defending dependent causation. However, reductive physicalists who opt for this strategy must acknowledge that they are thereby giving up on (Exclusion). This implies that they must concede victory to non-

reductive physicalism because non-reductive physicalism retains (Distinctness), whereas their position does not.

To sum up, we saw that expanding the physical domain to include disjunctions of physical properties saves reductive physicalism from the threat of epiphenomenalism only if (18) is true:

> (18) *M* is a disjunctive property with all of its disjuncts being properties in the physical domain, and *M* is sparse.

However, I have demonstrated from (24) to (28) that (18) is false because no disjunctive property is sparse.

3.3.3. Denying (10): Causally Efficacious Non-Sparse Properties

While discussing (14), we saw Kim concede that functional properties have no causal powers that go beyond the causal powers of their realizers. This implies that functional properties are not sparse, because "distinct [sparse] properties must represent distinct causal powers" (Kim, 1998, 103). However, in various places, he advances the strange position that functional properties are causally efficacious nonetheless:

> Functional properties, as second-order properties, do not bring new causal powers into the world: they do not have causal powers that go beyond the causal powers of their first-order realizers. According to the causal inheritance principle, the causal powers of an instance of a second-order property are identical with (or a subset of) the causal powers of the first-order realizer that is instantiated on that occasion. This means that second-order properties represent heterogeneous causal powers, but none that go beyond the causal powers of the first-order properties already in our domain over which they are defined. There

> are therefore no special problems about the causal powers of functional properties. And if any mental properties turn out to be functional properties, there are no special problems about their causal roles either. (...) According to the view being urged here, functional mental properties turn out, on account of their multiple realization, to be causally heterogeneous but not causally impotent. This solves the problem of causal efficacy for functionalizable mental properties. (Kim, 1998, 115-116)[50]

If what Kim says in this passage is true, then (10) is false:

> (10) If P is not a sparse property, then P is causally inefficacious.

So let's examine how Kim's argument might be refuted.

Kim's argument for the causal efficacy of functional properties in the cited passage is slightly dense and contains a number of substantive premises. Therefore, to evaluate it properly, I propose a more conspicuous formulation of the argument. First, we have the causal inheritance principle:

> (29) (Causal Inheritance principle)[51] If a second-order property F is instantiated on a given occasion in virtue of

[50] A similar claim is made in Kim (1997, 295-296)

[51] As an exegetical aside, some philosophers understood this principle to be claiming that in every case "a functionally realized feature inherits all of the token powers of its realizing feature" (Wilson, 2015, 369). and criticized the principle on such grounds. Hence, Wilson complains:

> Where a functional role may be played by multiple realizers, however, there is a case to be made that a functionally realized feature has, on a given occasion, only a proper subset of the

> one of its realizers, *R*, being instantiated, then the causal powers of this instance of *F* are identical with (or are a subset of) the causal powers of this instance of *R*. (Kim, 1998, 54)

Now, it follows directly from the definition of a second-order property that:

> (30) It is *always* the case that a second-order property *F* is instantiated on a given occasion in virtue of one of its realizers, *R*, being instantiated.

(29) and (30) collectively imply that:

> (31) *Every* instance of a second-order property *F* is causally efficacious, and *F*-instances have heterogeneous causal powers because the physical realizers of second-order properties, the *R*'s, are causally heterogeneous.

Here, however, Kim makes a curious leap from the causal efficacy of *F*-instances to the causal efficacy of *F*, claiming that "second-order properties represent heterogeneous causal powers". Since he does not provide any justification for this move, I can only speculate what he had in mind. It is plausible that he was assuming a principle like (32):

> (32) For any property *P*, if every *P*-instance is causally efficacious, then *P* is causally efficacious.

Then, (31) and (32) collectively entail:

> token powers of the feature realizing it on that occasion. (Wilson, 2015, 369)

However, as we can see in the cited passages, Kim never neglected nor excluded this possibility.

(33) A second-order property *F* is causally efficacious, and *F* represents heterogeneous causal powers.

The premises of the argument are (29), (30), and (32). (29) is obviously true for reductive physicalists because they will claim that *F*-instances are identical with *R*-instances. In this case, there is a trivial sense (identity) in which the *F*-instance "inherits" the causal powers of the *R*-instance. (30) is also uncontroversial since it follows directly from the fact that a second-order property is defined by existential quantification.

However, (32) is false. In chapter 1, we observed that anomalous monism succeeds in securing the causal efficacy of mental events. Yet, we found the view unsatisfactory because it suggests that a mental event's causal relations are fully and exclusively determined by its physical properties, making its mental properties causally irrelevant. Because of this, Kim and other critics categorized anomalous monism as a form of mental type-epiphenomenalism. However, if (32) were true, it would imply that mental properties are causally efficacious under anomalous monism. This is because every instance of mental properties is causally efficacious, even under this view. This consideration shows that the causal efficacy of *P*-instances is insufficient to guarantee the causal efficacy of *P*. Consequently, in proposing the necessary and sufficient condition for property causal efficacy, we added the requirement that *P*-instances must cause other property instantiations *in virtue of being an instance of P* for *P* to be causally efficacious:

(Property Efficacy) a property *P* is causally efficacious if and only if instantiations of *P* can, and do, cause other properties to be instantiated *in virtue of being an instance of P*.

According to (Property Efficacy), mental properties are not causally efficacious under anomalous monism because, according to this view, instantiations of mental events cause other properties to be instantiated *solely in virtue of an instance of a physical property.*

This consideration demonstrates that Kim's original argument for the causal efficacy of non-sparse properties, such as second-order properties, is unsound. Still, one might wonder whether it is possible for non-sparse properties to be causally efficacious according to (Property Efficacy). To examine this issue, suppose that *f*, an instance of the second-order property *F*, is identical with *r*, an instance of *F*'s realizer *R*. Because *f* = *r*, there is a straightforward sense in which *f*'s causal power is inherited from *r*'s causal power. But can *f* cause anything *in virtue of being an instance of F*? The answer depends on one's preferred conception of causation. If one endorses the nomological conception of causation, one might argue following Fodor (1989) that *f*'s causal relations are underwritten by various *ceteris paribus* laws. And since *F* will occur in such laws, there is a good sense in which *f* causes other events *in virtue of being an instance of F*. A similar story can be told regarding the counterfactual conception of causation because evaluations of counterfactual conditionals depend on laws. However, if one is working under the productive conception of causation, there is no such story to be told. According to this view, a cause must do some independent causal work in the production of the effect. However, all the powers that allow *f* (or *r*) to play such distinct and distinctive causal roles are given by *R*. *F*, on the other hand, does not bestow any causal power on *f* (or *r*) because it "do[es] not bring new causal powers into the world" (Kim, 1998, 115). Therefore, there is no good sense in which *f* causally produces another event *in virtue of being an instance of F*. This consideration shows that, under the productive conception of causation, in order for a property to count as causally efficacious, it must be sparse, bringing new causal powers into this world.

Indeed, given that one of the primary functions of sparse properties is to track or carve out causal powers, the concept of a "causally efficacious non-sparse property" itself appears contradictory. While proponents of the nomological or counterfactual conception of causation can make sense of *f* causing other events *in virtue of being an instance of F*, they argue that second-order properties like *F* are *sparse* in the first place, by invoking the proper subset account of realization.[52] I therefore conclude that causal efficacy cannot be separated from sparsity.

[52] This issue was discussed in section 3.1 in relation to premise (14).

Chapter 4. Conclusion

Let us take stock and summarize our discussion so far. We saw that the exclusion argument demonstrates the mutual inconsistency of the five theses, namely (Strong Supervenience), (Distinctness), (Causal Efficacy), (Exclusion), and (Closure). Kim used this result as an argument *against* non-reductive physicalism since he claimed that non-reductive physicalism is committed to all five premises. In addition, Kim argued that the exclusion argument establishes the position called "conditional physical reductionism", as physicalists have no other option but to deny either (Distinctness) or (Causal Efficacy):

> The position we have arrived at may be called *conditional physical reductionism*: the thesis that if mental properties are to be causally efficacious, they must be physically reducible. That is, to save mental causation we must reduce mentality. (Kim, 2005, 5)

Conditional physical reductionism, if true, would be a very strong argument for reductive physicalism because mental causation is non-negotiable for most philosophers.

However, we saw that the exclusion argument fails to establish conditional physical reductionism because non-reductive physicalists are *not* necessarily committed to all five premises of the exclusion argument. Instead, many non-reductive physicalists take the compatibilist approach of denying (Exclusion). However, this strategy comes at a cost because it requires a commitment to dependent causation, where the cause does not do any independent causal work in producing the effect. Arguably, this is a huge loss because dependent causation is incompatible with the productive conception of causation:

> (The Productive/Generative Conception of Causation) A cause is something that produces, or generates, or brings about its effects, something from which the effects *derive* their existence or occurrence. (Kim, 2007, 235)

Of course, compatibilists will argue that forfeiting the productive conception of causation is not a "cost" at all because they believe that the productive conception of causation is false in the first place, and that a better conception of causation capable of accommodating dependent causation, such as the nomological or counterfactual conception of causation, can be provided.

Without taking sides on this issue, we raised another question that would be of significant interest to Kim and other reductive physicalists. Assuming the truth of the productive conception of causation and (Exclusion), is conditional physical reductionism true? Kim's answer to this question would be a definite yes. This position may be called "weak conditional physical reductionism": the thesis that if mental properties are to be *productively* causally efficacious, they must be physically reducible.

In opposition to Kim, my main argument was that the exclusion argument does not even establish weak conditional physical reductionism. This is because, according to the productive conception of causation, denying (Distinctness) leads to the violation of (Causal Efficacy). To put it another way, under the productive conception of causation, reductive physicalism is a version of epiphenomenalism, rather than its rival. This implies that even if mental properties can be reduced to or identified with physical properties, they cannot be identified with any physical property that confers independent causal powers to its bearers. Instead, mental properties can only be reduced to those physical properties that are mere disjunctions or existential quantifications of causally efficacious physical properties. However, according to the productive conception of causation, disjunctions and

existential quantifications are causally irrelevant because they do not bring any new causal powers into the world.

In my argument against weak conditional physical reductionism, I made reference to various points already noted by other philosophers. For example, the crucial argument for (13) consisted of the fact that most macro-properties are disjunctions, rather than conjunctions, of basic physical properties. This idea has already been stated by various philosophers such as Loewer (2002), Block (2003), and Schaffer (2004). Furthermore, in defending (9), I argued against Kim's halfway eliminative approaches to mentality, such as the local reduction strategy and the token-identity strategy. Here, Marras (2003) has already made the objection that local reduction is not substantially different from the token-identity theory, and Walter (2008) has argued that mere token-identity is too weak to fulfill the aspirations of reductive physicalism.

> (13) *M* is not a property that is micro-based on entities and properties in the physical domain. In other words, *M* is not a physical micro-based property.
>
> (9) A mental property *M* is identical with a physical property *P*.

However, these considerations were usually raised as standalone objections with the purpose of defending non-reductive physicalism from Kim's criticisms. Nobody saw that such points could be used to turn the tables on Kim's position. Until now, all parties to the debate agreed that, given (Exclusion), the exclusion argument does indeed force us to embrace reductive physicalism. In other words, no one doubted the truth of weak conditional physical reductionism. That is why so many non-reductive physicalist strategies focused on criticizing (Exclusion). My original contribution to the debate is that I depart from this convention by using various points already mentioned

by other philosophers to argue that the exclusion argument does not support reductive physicalism, *even if* we grant the truth of (Exclusion).

If my argument is sound, it carries several implications. First, anyone committed to the productive conception of causation must look outside physicalism to save mental causation. Non-reductive physicalism is not an option for her since anyone committed to the productive conception of causation must accept the truth of (Exclusion). Neither can she deny (Distinctness) and endorse reductive physicalism because, according to her view of causation, it is merely a form of epiphenomenalism. Therefore, she must deny at least one of (Strong Supervenience) and (Closure) to retain (Causal Efficacy). This leads to anti-physicalism because the two premises are necessary commitments of physicalism. Second, reductive physicalists find themselves in a dilemma. They must either concede that their position leads to epiphenomenalism or follow non-reductive physicalists in rejecting the productive conception of causation and (Exclusion). The first horn of the dilemma is obviously unacceptable. However, the second horn is not much better as it entails that non-reductive physicalism is a superior solution to the exclusion argument than reductive physicalism. This is because non-reductive physicalism only denies (Exclusion), whereas reductive physicalism denies both (Exclusion) and (Distinctness). Therefore, my argument demonstrates that *the exclusion argument is in fact an argument against reductive physicalism*. This conclusion is surprising and at the same time ironic given that its inventor, Kim, intended exactly the opposite.

Bibliography

Aimar, S. (2011), "Counterfactuals, Overdetermination and Mental Causation", *Proceedings of the Aristotelian Society* 111 (3pt3): 469–77.

Alexander, S. (1920), *Space, Time, and Deity*, Macmillan.

Anscombe, E. (1993), "Causality and Determination", in E. Sosa M. Tooley (eds.), *Causation*. Oxford University Press, pp. 88–104.

Armstrong, D. (1997), *A World of States of Affairs*, Cambridge University Press.

Baker, L. (1993), "Metaphysics and Mental Causation", in J. Heil & A. Mele (eds.), *Mental Causation*. Oxford University Press, pp. 75–96.

Bennett, K. (2003), "Why the Exclusion Problem Seems Intractable and How, Just Maybe, to Tract It", *Noûs* 37 (3): 471–97.

----- (2008), "Exclusion Again", in J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press, pp. 280–307.

Bernstein, S. (2016), "Overdetermination Underdetermined", *Erkenntnis* 81 (1):17–40.

Block, N. (1997), "Anti-Reductionism Slaps Back", *Noûs* 31 (s11):107–32.

----- (2003), "Do Causal Powers Drain Away?", *Philosophy and Phenomenological Research* 67 (1):133–50.

Crane, T. and Mellor, D. (1990), "There is No Question of Physicalism", *Mind* 99 (394):185–206.

Crane, T. and Árnadóttir, S. (2013), "There is No Exclusion Problem", in E. Lowe, S. Gibb, and R. Ingthorsson (eds.), *Mental Causation and Ontology.* Oxford University Press, pp. 248-66.

Crisp, T. and Warfield, T. (2001), "Kim's Master Argument", *Noûs* 35 (2): 304-16.

Davidson, D. (1970), "Mental Events", in L. Foster & J. Swanson (eds.), *Experience and Theory*. Clarendon Press, pp. 207-24.

Dowe, P. (2000), *Physical Causation*, Cambridge University Press.

Feigl, H. (1958), "The 'Mental' and the 'Physical'", *Minnesota Studies in the Philosophy of Science* 2: 370-497.

Fodor, J. (1989), "Making Mind Matter More", *Philosophical Topics* 17 (1): 59-79.

----- (1997), "Special Sciences: Still Autonomous after All these Years", *Noûs* 31 (S11): 149-63.

Horgan, T. (1993), "From Supervenience to Superdupervenience: Meeting the Demands of a Material World", *Mind* 102 (408): 555-86.

Jackson, F. (1998), *From Metaphysics to Ethics: A Defense of Conceptual Analysis*, Oxford, GB: Oxford University Press.

Kim, J. (1984), "Epiphenomenal and Supervenient Causation", *Midwest Studies in Philosophy* 9 (1): 257-70.

----- (1992), "Multiple Realization and the Metaphysics of Reduction", *Philosophy and Phenomenological Research* 52 (1): 1-26.

----- (1997), "Does the Problem of Mental Causation Generalize?", *Proceedings of the Aristotelian Society* 97 (3): 281-97.

----- (1998), *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*, MIT Press.

----- (1999), "Making Sense of Emergence", *Philosophical Studies* 95 (1-2): 3-36.

----- (2005), *Physicalism, or Something Near Enough*, Princeton University Press.

----- (2006), "Being Realistic About Emergence", in P. Clayton & P. Davies (eds.), *The Re-Emergence of Emergence*. Oxford University Press, pp. 189-202.

----- (2007), "Causation and Mental Causation", in B. McLaughlin & J. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*. Blackwell, pp. 227-42.

----- (2009), "Mental Causation", in B. Ansgar, B. McLaughlin, and S. Walter (eds.), *The Oxford Handbook of Philosophy of Mind*. Oxford University Press, pp. 29-52.

----- (2010), *Philosophy of Mind*, Boulder: Westview Press.

----- (2011), "From Naturalism to Physicalism: Supervenience Redux", *Proceedings and Addresses of the American Philosophical Association* 85, no. 2: 109 - 34.

Lewis, D. (1973), "Causation", *Journal of Philosophy* 70 (17): 556 - 67.

----- (1983), "New Work for a Theory of Universals", Australasian Journal of Philosophy 61 (4): 343-77.

Lowe, E. (2003), "Physical Causal Closure and the Invisibility of Mental Causation", in S. Walter & H. Heckmann (eds.), *Physicalism and Mental Causation*. Imprint Academic. pp. 137-54.

Loewer, B. (2002), "Comments on Jaegwon Kim's *Mind and the Physical World*", *Philosophy and Phenomenological Research* 65 (3):655－662.

----- (2007), "Mental Causation, or Something Near Enough", in B. McLaughlin & J. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*. Blackwell, pp. 243-64.

Marras, A. (2002), "Kim on Reduction", *Erkenntnis* 57 (2):231-257.

McLaughlin, B. (1989), "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical", *Philosophical Perspectives* 3: 109-36.

Melnyk, A. (2003), *A Physicalist Manifesto*, New York: Cambridge University Press.

Nagel, E. (1961), *The Structure of Science. Problems in the Logic of Explanation,* New York: Harcourt, Brace & World, Inc.

O'Connor, T. and Churchill, R. (2010), "Is Non-Reductive Physicalism Viable Within a Causal Powers Metaphysic?", in G. Macdonald and C. Macdonald (eds.), *Emergence in Mind*. Oxford University Press, pp. 43-60.

O'Connor, T. and Wong, H. (2005), "The Metaphysics of Emergence", *Noûs* 39 (4): 658－78.

Papineau, D. (2001), "The Rise of Physicalism", in C. Gillett and B. Loewer (eds.), *Physicalism and Its Discontents*. Cambridge University Press, pp. 3－36.

Putnam, H. (1975), "The Nature of Mental States", in H. Putnam, *Mind, Language and Reality: Philosophical Papers*. Cambridge University Press, pp. 429-40.

Salmon, W. (1994), "Causality Without Counterfactuals", *Philosophy of Science* 61: 297-312.

Schaffer, J. (2003), "Overdetermining Causes", Philosophical Studies 114 (1-2): 23-45.

----- (2004), "Two Conceptions of Sparse Properties", *Pacific Philosophical Quarterly* 85 (1): 92-102.

Shoemaker, S. (2001), "Realization and Mental Causation", in C. Gillett and B. Loewer (eds.), *Physicalism and Its Discontents*. Cambridge University Press, pp. 74-98.

Sider, T. (2003), "What's So Bad about Overdetermination?", *Philosophy and Phenomenological Research* 67 (3): 719-26.

Smart, J. (1959), "Sensations and Brain Processes", *Philosophical Review* 68 (April): 141-56.

Stoljar, D. (2001), "Physicalism", *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.), URL = 〈https://plato.stanford.edu/entries/physicalism/〉.

Van Gulick, R. (1992), "Three Bad Arguments for Intentional Property Epiphenomenalism", *Erkenntnis* 36 (3): 311-31.

Walter, S. (2008), "The Supervenience Argument, Overdetermination, and Causal Drainage: Aseessing Kim's Master Argument", *Philosophical Psychology* 21 (5): 673-96.

Wilson, J. (2005), "Supervenience-based Formulations of Physicalism", *Noûs* 39 (3):426-59.

----- (2015), "Metaphysical Emergence: Weak and Strong", in T. Bigaj & C. Wüthrich (eds.), *Metaphysics in Contemporary Physics*. Poznan Studies in the Philosophy of the Sciences and the Humanities, pp. 251-306.

Won, C. (2014), "Overdetermination, Counterfactuals, and Mental Causation", *Philosophical Review* 123 (2): 205-29.

국 문 초 록

인과적 배제 논증에 의하면, 물리적으로 환원 불가능한 심적 사건을 원인으로 갖는 임의의 물리적 사건은 다른 물리적 사건에 의해 인과적으로 과잉결정 된다. 김재권은 이와 같은 체계적 심-물 (mind-body) 인과적 과잉결정을 받아들일 이유가 없다고 주장한다. 그렇다면 소위 "물리적으로 환원 불가능한 심적 원인"은 실제로는 물리적으로 환원 가능하거나 혹은 인과적 효력을 갖지 않아야 한다. 달리 말해, 인과적 배제 논증의 교훈은 심적 인과를 구제하기 위해선 정신을 물리적으로 환원해야 한다는 것이다. 김재권은 이러한 입장을 "조건적 물리적 환원주의"(conditional physical reductionism)라 명명한다. 조건적 물리적 환원주의는 정신이 물리적으로 환원된다고 주장하지는 않기 때문에, 환원적 물리주의보다는 약한 입장이다. 그러나 대다수의 철학자들이 정신의 인과적 효력을 포기하길 원하지 않는다는 점을 고려할 때, 조건적 물리적 환원주의는 환원적 물리주의를 강력하게 지지한다.

하지만 다수의 비환원적 물리주의자들은 인과적 배제 논증이 조건적 물리적 환원주의를 입증하지 않는다고 반론한다. 더 구체적으로, 그들은 인과적 양립가능주의(causal compatibilism)를 인과적 배제 논증에 대한 해결책으로서 지지한다. 이 관점에 따르면, 원인들 사이에 밀접한 양상적 연결이 성립할 경우에, 결과는 둘 이상의 충분 원인을 가질 수 있다. 그런데 체계적 심-물 인과의 사례에서 심적 원인은 물리적 원인에 수반한다. 그러므로 체계적 심-물 인과는 전혀 문제적이지 않다는 것이 인과적 양립가능주의의 입장이다.

이에 대해 김재권은 인과에 대한 "두꺼운" 이론들 중 하나인 인과에 대한 생산적 관점(the productive conception of causation)이 체계적 심-물(mind-body) 인과적 과잉결정을 거부하기 위해 필요하다고 인정한 후, 인과에 대한 생산적 관점을 옹호하는 방식으로 양립가능주의에 대응한다. 인과에 대한 생산적 관점에 따르면, 원인은 그 결과들을 생산하는 것이고, 결과들이 그로부터 자신들의 존재를 이끌어내는 (derive from) 것이다.

김재권의 노력에도 불구하고, 인과에 대한 생산적 관점이 타당한 지 여부는 여전히 논쟁적이다. 하지만, 김재권과 인과적 양립가능주의자들은 체계적 심-물 인과적 과잉결정이 인과에 대한

생산적 관점과 양립 불가능하다는 점에 있어서는 합의를 보인다. 이러한 점을 고려할 때, *인과에 대한 생산적 관점의 참을 전제할 경우*, 인과적 배제 논증은 조건적 물리적 환원주의를 입증하는 것처럼 보인다. 실제로, 필자가 아는 한, 그 어떤 철학자도 이러한 귀결이 성립하는 지 여부를 문제삼지 않는다.

본 논문의 주요 목적은 그러한 귀결이 성립하지 않음을 보이는 것이다. 즉, *인과에 대한 생산적 관점이 참일 지라도*, 인과적 배제 논증은 조건적 물리적 환원주의를 입증하지 못한다. 이러한 결론에 도달하기 위한 필자의 기본적인 전략은, 인과에 대한 생산적 관점 하에서, 심적 속성들은 *오로지* 아무런 인과적 효력도 갖지 않는 물리적 속성들로만 환원 가능하다고 주장하는 것이다.

필자의 논변이 타당하다면, 김재권을 비롯한 환원적 물리주의자들은 인과에 대한 생산적 관점을 지지할 지 여부에 대한 딜레마에 직면하게 된다. 만약 김재권이 인과에 대한 생산적 관점을 받아들인다면, 그는 환원적 물리주의가 사실상 부수현상론의 한 형태에 불과하다는 점을 인정해야 한다. 하지만 김재권이 심적 인과의 실재성을 타협 불가능한 것으로 여긴다는 점으로 미루어 볼 때, 이러한 결론은 수용 불가능하다. 다른 한편으로, 만약 김재권이 인과에 대한 생산적 관점을 거부한다면, 그는 체계적 심-물 인과적 과잉결정을 수용해야 한다. 그러나 이 경우에는 심적 인과를 구제하기 위해 정신을 물리적으로 환원할 필요가 사라져버린다. 체계적 심-물 인과적 과잉결정이 수용된 이상, 인과적 양립가능주의의 전략만으로도 정신의 인과적 효력을 보장할 수 있기 때문이다. 이를 통해 비환원적 물리주의가 환원적 물리주의보다 인과적 배제 논증에 대해 더 효과적인 해결책을 제시한다는 결론이 도출된다. 이처럼 김재권이 딜레마의 어느 한 쪽 뿔도 수용할 수 없다는 점을 고려할 때, 인과적 배제 논증은 환원적 물리주의를 옹호하는 것이 아니라 오히려 그것에 반대하는 논증임이 드러난다.

주요어 : 인과적 배제 논증, 김재권, 조건적 물리적 환원주의, 환원적 물리주의, 인과적 양립가능주의, 인과에 대한 생산적 관점
학 번 : 2020-28074